The phase transition in random graphs and random graph processes

DISSERTATION

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften (doctor rerum naturalium)
im Fach Informatik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät II
der Humboldt-Universität zu Berlin

von
Herr cand.scient. Taral Guldahl Seierstad
geboren am 5. September 1979 in Lillehammer

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Christoph Markschies

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II:
Prof. Dr. Wolfgang Coy

Gutachter:

1. Prof. Dr. Hans Jürgen Prömel
2. Prof. Dr. Tomasz Łuczak
3. Prof. Dr. Stefan Felsner

eingereicht am:	29. Januar 2007
Tag der mündlichen Prüfung:	28. Juni 2007

Zusammenfassung

Zufallsgraphen sind Graphen, die durch einen zufälligen Prozess erzeugt werden. Ein im Zusammenhang mit Zufallsgraphen häufig auftretendes Phänomen ist, dass sich die typischen Eigenschaften eines Graphen durch Hinzufügen einer relativ kleinen Anzahl von zufälligen Kanten radikal verändern. Dieses Phänomen wurde zuerst in den bahnbrechenden Arbeiten von Erdős und Rényi untersucht.

Wir betrachten den Zufallsgraphen $G(n,p)$, der n Knoten enthält und in dem zwei Knoten unabhängig und mit Wahrscheinlichkeit p durch eine Kante verbunden sind. Erdős und Rényi zeigten, dass ein Graph für $p = \frac{c}{n}$ und $c < 1$ mit hoher Wahrscheinlichkeit aus Komponenten mit $O(\log n)$ Knoten besteht. Für $p = \frac{c}{n}$ und $c > 1$ enthält $G(n,p)$ mit hoher Wahrscheinlichkeit genau eine Komponente mit $\Theta(n)$ Knoten, welche viel größer als alle anderen Komponenten ist.

Der Punkt in der Entwicklung des Graphen, an dem sich die Komponentenstruktur durch eine kleine Erhöhung der Anzahl von Kanten stark verändert, wird *Phasenübergang* genannt. Im $G(n,p)$ passiert er bei $p = \frac{1}{n}$. Darüber hinaus durchlebt $G(n,p)$ einen sogenannten Doppelsprung. Wenn die Wahrscheinlichkeit p von $\frac{1-\varepsilon}{n}$ auf $\frac{1}{n}$ steigt, dann wächst die größte Komponente von $O(\log n)$ auf $\Theta(n^{2/3})$ Knoten. Ist schließlich p gleich $\frac{1+\varepsilon}{n}$, dann besteht die größte Komponente aus $\Theta(n)$ Knoten. Wenn $p = \frac{1+\varepsilon}{n}$, wobei $\varepsilon = \varepsilon(n)$ eine Funktion von n ist, die gegen 0 geht, sind wir in der *kritischen Phase*, welche eine der interessantesten Phasen der Entwicklung des Zufallsgraphen ist. In diesem Fall hängt die Komponentenstruktur des Graphen von der Geschwindigkeit ab, mit welcher ε gegen 0 konvergiert.

In dieser Arbeit betrachten wir drei verschiedene Modelle von Zufallsgraphen. In Kapitel 4 studieren wir den Minimalgrad-Graphenprozess. In diesem Prozess werden sukzessive Kanten vw hinzugefügt, wobei v ein zufällig ausgewählter Knoten von minimalem Grad ist. Wir beweisen, dass es in diesem Graphenprozess einen Phasenübergang und wie im $G(n,p)$ einen Doppelsprung gibt.

Die zwei anderen Modelle sind Zufallsgraphen mit einer vorgeschriebenen Gradfolge und zufällige gerichtete Graphen. Für diese Modelle wurde bereits in den Arbeiten von Molloy und Reed (1995), Karp (1990) und Łuczak (1990) gezeigt, dass es einen Phasenübergang bezüglich der Komponentenstruktur gibt. In dieser Arbeit untersuchen wir in Kapitel 5 und 6 die kritische Phase dieser Prozesse genauer und zeigen, dass sich diese Modelle ähnlich zum $G(n,p)$ verhalten.

Schlagwörter:
Zufallsgraphen, Phasenübergang, kritische Phase, größte Komponente

Abstract

Random graphs are graphs which are created by a random process. They are used among other places in the study of large networks, and in the analysis of the performance of algorithms.

A common phenomenon in random graphs is that the typical properties of a graph change radically by the addition of a relatively small number of random edges. This phenomenon was first investigated in the seminal papers of Erdős and Rényi.

We consider the graph $G(n,p)$ which contains n vertices, and where any two vertices are connected by an edge independently with probability p. Erdős and Rényi showed that if $p = \frac{c}{n}$ and $c < 1$, then with high probability $G(n,p)$ consists of components with $O(\log n)$ vertices. If $p = \frac{c}{n}$ and $c > 1$, then with high probability $G(n,p)$ contains exactly one component, called the *giant component*, with $\Theta(n)$ vertices, which is much larger than all other components.

The *phase transition* in a random graph refers to the point at which the giant component is formed. In $G(n,p)$ this is when $p = \frac{1}{n}$. Moreover, $G(n,p)$ undergoes a so-called *double jump* at this point. When the probability p increases from $\frac{1-\varepsilon}{n}$ to $\frac{1}{n}$, the largest component grows from $O(\log n)$ to $\Theta(n^{2/3})$ vertices. When p becomes $\frac{1+\varepsilon}{n}$, the graph contains a giant component with $\Theta(n)$ vertices. If we let $p = \frac{1+\varepsilon}{n}$, where ε is a function of n tending to 0, we are in the *critical phase* of the random graph, which is one of the most interesting phases in the evolution of the random graph. In this case the structure depends on how fast ε tends to 0.

In this dissertation we consider three different random graph models. In Chapter 4 we consider the so-called minimum degree graph process. In this process edges vw are added successively, where v is a randomly chosen vertex with minimum degree. We prove that a phase transition occurs in this graph process as well, and also that it undergoes a double jump, similar to $G(n,p)$.

The two other models we will consider, are random graphs with a given degree sequence and random directed graphs. In these models the point of the phase transition has already been found, by Molloy and Reed (1995), Karp (1990) and Łuczak (1990). In Chapter 5 and 6 we investigate the critical phase of these processes, and show that their behaviour resembles $G(n,p)$.

Keywords:
random graphs, phase transition, critical phase, giant component

Acknowledgement

I would like to thank my supervisor Prof. Dr. Hans Jürgen Prömel for letting me join his research group at the Humboldt University, and the European Graduate Program "Combinatorics, Geometry, and Computation" for scholarship and support. I also want to thank the members of the research group for algorithms and complexity at the Humboldt University for interesting discussions and seminars, especially Prof. Dr. Anusch Taraz who helped me much during my first year, and Dr. Mihyun Kang with whom I have had many stimulating discussions.

Furthermore I want to thank Dr. Manuel Bodirsky who gave me feedback on a draft of this dissertation. I am also grateful to Prof. Dr. Michał Karoński and the members of the Department for Discrete Mathematics at the Adam Mickiewicz University for their hospitality, and in particular to Prof. Dr. Tomasz Łuczak for very inspiring collaboration, and to Dr. Małgorzata Bednarska who was very helpful during my six-month stay in Poznań.

Finally I want to thank my family for their support and encouragement through many years.

Preface

In this thesis we study the evolution of random graph processes. The study of random graphs was initiated by Erdős and Rényi around 1960, and became a flourishing research area in the following decades. We will mostly concern ourselves with a phenomenon which occurs in several random graph processes, generally known as the *phase transition*, namely that the component structure of a random graph changes substantially caused by the addition of relatively few random edges.

One of the motivations for studying random graphs is the desire to describe a "typical" graph. For example, if we consider all labelled graphs on n vertices, it is known that the vast majority of the graphs are connected, contain a copy of any fixed graph F, and has chromatic number close to $\frac{n}{2\log_2 n}$, provided that n is large enough. Moreover, the proportion of graphs not having these properties decreases as n grows. We therefore feel justified in saying that "almost all" graphs have these properties. In the terminology of random graphs we say that a random graph has these properties with probability tending to 1 as the number of vertices tends to infinity.

More interesting results can be obtained if we restrict ourselves to subclasses of graphs, for example by fixing the number of edges and asking what the typical properties of a graph with n vertices and m edges are. Often we think of random graphs as states in a process. We begin at time 0 with an empty graph on n vertices. Then as the time goes, we add edges to the graphs at random, either uniformly or according to some other random procedure. A discovery of Erdős and Rényi was that many graph properties enjoy so-called *threshold phenomena*: when the number of edges in the random graph is significantly smaller than the threshold, it has the property with probability very close to 0, while if the number of edges is significantly greater than the threshold, it has the property with probability very close to 1.

The main topic of this dissertation is the phenomenon known as the phase transition. A random graph with n vertices and $0.49n$ edges is very likely to consist of many small components, none of which has more than

$O(\log n)$ vertices, while a random graph with n vertices and $0.51n$ edges most probably contains a unique large component containing a linear number of vertices. This large component is called the *giant component*, and there has been much interest in studying the evolution of this component, and in particular in examining the particular point in a graph process where the giant component is first formed. This point of a graph process is generally referred to as the *phase transition*, because of the similarities to the physical phenomenon of substances turning from one phase to another by a small change in temperature or pressure.

The phenomenon is also related to *percolations* in statistical mechanics. In percolation theory a typical question is whether, and with which probability, the centre of a porous stone becomes wet if the stone is put into water. The stone can be represented by a graph, with different points being adjacent if they are connected by a hole in the stone.

The thesis is organised as follows. Chapter 1 contains an introduction of random graph theory and presents the random graph models we will examine later in the thesis. In Chapter 2 we present some well-known results from various areas of mathematics, which will be used in the analysis of the random graph processes later on. In Chapter 3 we discuss recursive trees and the early phase of the minimum degree graph process, while Chapter 4 is concerned with the phase transition of that process. In Chapter 5 we study the critical phase of random graphs with a given degree sequence, and in Chapter 6 we consider the critical phase for random digraphs.

Chapters 4 and 5 are joint work with Mihyun Kang, and are based on [29] and [28] respectively. Chapter 6 is joint work with Tomasz Łuczak and is based on [40].

Contents

Chapter 1

Introduction

The starting point of the theory of random graphs is generally considered to be a series of papers by Erdős and Rényi from between 1959 and 1961, in which random graphs were studied in their own right for the first time. Two notable papers are [17] from 1959 and [16] from 1960; some of the results contained therein will be given below. The field is vast, so we will mostly restrict ourselves to topics which are relevant to later chapters.

1.1 The basic random graph models

The model which Erdős and Rényi first concentrated on is the model we now know as the $G(n,m)$ model. Let $\Omega_{n,m}$ be the set of those graphs on the vertex set $[n] = \{1,\dots,n\}$ which contain exactly m edges. Then $G(n,m)$ is a graph chosen uniformly at random from $\Omega_{n,m}$. Thus, if G is a graph with n vertices and m edges, then

$$\mathbb{P}\left[G(n,m) = G\right] = \binom{\binom{n}{2}}{m}^{-1}.$$

Another model, which was also described in [16], but was first introduced by Gilbert [21], is the $G(n,p)$ model. We let Ω_n be the set of graphs on n vertices, and we let $G(n,p)$ be chosen at random from Ω_n such that if G has n vertices and m edges, then

$$\mathbb{P}\left[G(n,p) = G\right] = p^m(1-p)^{\binom{n}{2}-m}.$$

Equivalently, every pair of vertices in $G(n,p)$ forms an edge with probability p, independently of every other pair of vertices. This model is called the *binomial model* of random graphs.

Typically p and m are not fixed numbers, but functions of n, and we are interested in the asymptotic properties of $G(n,p)$ or $G(n,m)$ as n tends to infinity. In general, if $\mathcal{P}$ is a graph property, we want to determine the limit of the probability that $G(n,p)$ or $G(n,m)$ has $\mathcal{P}$ as n tends to infinity. Often this probability tends to either 0 or 1; we say that $G(n,p)$ (or $G(n,m)$) has the property $\mathcal{P}$ *asymptotically almost surely*, abbreviated a.a.s., if the probability that $G(n,p)$ (or $G(n,m)$) has $\mathcal{P}$ tends to 1 as n tends to infinity.

The two models $G(n,m)$ and $G(n,p)$ have very similar properties when $m \sim \binom{n}{2}p$. In this case the expected number of edges in $G(n,p)$ is about m. Although the probability that $G(n,p)$ actually contains exactly m edges is generally very small, in many cases theorems proved for $G(n,p)$ also hold for $G(n,m)$ with $m \sim \binom{n}{2}p$, and vice versa; see Łuczak [36]. We will mostly consider $G(n,p)$, rather than $G(n,m)$, since the fact that the edges in $G(n,p)$ are present independently of each other makes it more comfortable to work with than $G(n,m)$, although some of the graph processes we will consider later have more in common with $G(n,m)$.

We will also consider random graphs from a dynamical viewpoint. In one such model we start with an empty graph G on n vertices which changes over time. At every step in the process we choose a pair of vertices $\{v,w\}$ uniformly at random from the set of pairs of nonadjacent vertices, and add the edge vw to the graph. After precisely m edges have been added, the probability distribution of G is the same as that of $G(n,m)$.

It is also possible to consider $G(n,p)$ as a dynamical graph process. Here we give every potential edge a "birth-time" chosen uniformly at random from the interval $[0,1]$. Then we let p increase gradually from 0 to 1. The graph $G(n,p)$ then consists of those edges which have birth-time at most p. Thus the graph process starts as an empty graph and grows until it becomes complete, at the latest when the time reaches 1.

1.2 Notation

Before we can go further, we have to introduce some notation, in particular to deal with the asymptotic behaviour of functions. Let $f(n)$ and $g(n)$ be two positive functions of n. We write $f(n) = O(g(n))$ if there is a constant C such that $f(n) \le Cg(n)$ for all (large enough) n, and we write $f(n) = \Omega(g(n))$ if there is a constant $c > 0$ such that $f(n) \ge cg(n)$ for all (large enough) n. If $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$, we write $f(n) = \Theta(g(n))$ or $f(n) \asymp g(n)$. If $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$, we write $f(n) = o(g(n))$, or $f(n) \ll g(n)$, and if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$, we write $f(n) \sim g(n)$.

As already mentioned, an event happens *asymptotically almost surely* (a.a.s.) if its probability tends to 1 as n tends to infinity. An event happens *almost surely* (a.s) if its probability equals 1.

If A is an event, I_A is the indicator variable of A, which is equal to 1 if A holds and 0 otherwise. A special case of an indicator variable is the Kronecker delta δ_{ij}, which equals 1 if $i = j$ and is 0 otherwise. The set $\{1, \ldots, n\}$ is denoted by $[n]$. If $\mathbf{v}$ is a vector, its transpose is denoted by $\mathbf{v}'$. All logarithms are natural.

1.3 Threshold functions

One of the important discoveries by Erdős and Rényi was that many properties exhibit a so-called *threshold* phenomenon: a small change in the number of edges of a random graph may have a big impact on the probability that the random graph has a certain property. Let $\mathcal{P}$ be a graph property. Then we say that $t(n)$ is a *threshold function* for $\mathcal{P}$ if

$$\lim_{n\to\infty} \mathbb{P}\left[G(n,p) \text{ has } \mathcal{P}\right] = \begin{cases} 0 & \text{if} \quad p(n) \ll t(n), \\ 1 & \text{if} \quad p(n) \gg t(n). \end{cases}$$

Bollobás and Thomason [13] proved that every property which is preserved by the addition of edges has such a threshold function. However, for several properties the increase in probability happens even more abruptly. A function $t(n)$ is said to be a *sharp threshold* for $\mathcal{P}$ if for every $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left[G(n,p) \text{ has } \mathcal{P}\right] = \begin{cases} 0 & \text{if} \quad p(n) < (1-\varepsilon)t(n), \\ 1 & \text{if} \quad p(n) > (1+\varepsilon)t(n). \end{cases}$$

In [17] Erdős and Rényi proved that the property of a graph being connected has the function $t(n) = \frac{\log n}{n}$ as a sharp threshold. In fact, they described the transition phase, in which the limit probability increases from 0 to 1, much more precisely. Let $p(n) = \frac{\log n + c(n)}{n}$. Then

$$\lim_{n\to\infty} \mathbb{P}\left[G(n,p) \text{ is connected}\right] = \begin{cases} 0 & \text{if} \quad c(n) \to -\infty, \\ e^{-e^{-c}} & \text{if} \quad c(n) \to c \in \mathbb{R}, \\ 1 & \text{if} \quad c(n) \to \infty. \end{cases}$$

They also proved that $G(n,m)$, when considered as a dynamic process, becomes connected a.a.s. at the moment when the last isolated vertex disappears. This holds even more generally: Erdős and Rényi [15] showed that $G(n,m)$ becomes k-connected a.a.s. at the moment when the last vertex

of degree $k-1$ disappears; this has a.a.s. happened when $m = \frac{n}{2}(\log n + k \log\log n + \alpha(n))$ where $\alpha(n) \to \infty$. The first Hamiltonian cycle appears in $G(n,m)$ a.a.s. when the graph becomes 2-connected, which is a.a.s. when the last vertex of degree 1 disappears. This was shown by Komlós and Szemerédi [33] and Bollobás [9].

All the properties just mentioned thus have sharp thresholds. A threshold which is not sharp is called a *coarse threshold*. An example of a property with a coarse threshold is that of subgraph containment. Let F be a fixed graph with v vertices and e edges. The *density* of F is defined to be $d(F) = \frac{e}{v}$, and the *maximal density* $m(F)$ is the density of the subgraph of F with the highest density. Bollobás [10] showed that

$$\lim_{n\to\infty} \mathbb{P}\left[G(n,p) \supset F\right] = \begin{cases} 0 & \text{if } p \ll n^{-1/m(F)}, \\ 1 & \text{if } p \gg n^{-1/m(F)}. \end{cases}$$

In the case of *balanced graphs* — that is graphs for which $m(F) = d(F)$ — this was proved already by Erdős and Rényi [16]. If $p \asymp n^{-1/m(F)}$, then $G(n,p)$ contains a copy of F with probability bounded away from 0 and 1.

More generally it appears that local properties often have coarse thresholds, while global properties often have sharp thresholds. We will now consider a property with a sharp threshold, namely the property that a random graph has a component of order $\Theta(n)$. The short period of time in which this component evolves is dubbed the *phase transition* and is arguably one of the best studied periods of the entire evolution of random graphs.

1.4 The phase transition

The *phase transition* refers to the sudden change in the component structure of many random graph processes. In the random graph $G(n,p)$ it happens around the time $\frac{1}{n}$. It was first described in [16] and has later been examined in minute detail by several authors. Let us consider $G(n,p)$ where $p = \frac{c}{n}$ for a constant c. When $c < 1$, the graph $G(n,p)$ consists a.a.s. of small components, all of which have $O(\log n)$ vertices. As we let c increase, these components merge and grow larger, and as soon as $c > 1$, there is a.a.s. a unique large component which consists of $\Theta(n)$ vertices. Thus, in the very short time from $c = 1-\varepsilon$ to $c = 1+\varepsilon$, for any $\varepsilon > 0$, many of the small components join to form one large component. This component is known as the *giant component*. If we stop the time exactly at the point when $c = 1$, we will see that the largest component has in the order of $n^{2/3}$ vertices, and that there are many components of roughly the same size. As soon as $c > 1$, however, there is only one large component; the second largest component

has $O(\log n)$ vertices. Thus the order of the largest component first makes a jump from $\Theta(\log n)$ vertices to $\Theta(n^{2/3})$, and then to $\Theta(n)$; for this reason the phenomenon is also called the *double jump*. We state this fundamental theorem here.

Theorem 1.1 (Erdős, Rényi 1960)**.** *Let c be a positive constant, and let $p = \frac{c}{n}$.*

(i) If $c < 1$, then a.a.s. no component in $G(n,p)$ contains more than one cycle, and no component has more than $\frac{\log n}{c-1-\log c}$ vertices.

(ii) If $c = 1$ and $\omega(n)$ is a function tending to infinity as $n \to \infty$, then $G(n,p)$ a.a.s. contains at least one component with more than $n^{2/3}/\omega(n)$ vertices and no component with more than $n^{2/3}\omega(n)$ vertices.

(iii) If $c > 1$, then $G(n,p)$ a.a.s. contains a component with $(d + o(1))n$ vertices, where $d + e^{-cd} = 1$, while every other component has at most $\frac{\log n}{c-1-\log c}$ vertices and contains at most one cycle.

We present a sketch of a proof of part (i) and (iii) of this theorem in Chapter 2.2.2, using branching processes. The original proof of this theorem, by Erdős and Rényi, uses a counting argument.

Theorem 1.1 tells us much about the random graph $G(n,p)$ when $c \neq 1$, but the most interesting question is arguably to find out how the giant component is formed, which happens when $c = 1$. This is called the *critical phase* of the graph process. In this period there are several large components of roughly the same size vying for dominance. As p increases, these large components eat many of the small components and merge with each other until a single giant component remains. One important problem is to determine at which point there is a component which a.a.s. remains largest until the end of the process in the dynamical model. It turns out that the appropriate parametrisation of p in the critical phase is

$$p = \frac{1}{n} + \frac{\lambda}{n^{4/3}}. \tag{1.1}$$

Bollobás [8] goes a long way to explain the development in this phase, but a fully satisfactory answer was only found by Łuczak [35] in 1990, thirty years after Erdős and Rényi first described the phase transition.

Theorem 1.2 (Łuczak 1990)**.** *Let $np = 1 + \lambda n^{-1/3}$, and let $L_k(G(n,p))$ be the order of the kth largest component in $G(n,p)$.*

(i) If $\lambda \to -\infty$, then a.a.s. $L_1(G(n,p)) \ll n^{2/3}$.

(ii) If $\lambda \to \infty$, then a.a.s. $L_1(G(n,p)) \gg n^{2/3} \gg L_2(G(n,p))$. Furthermore $L_1(G(n,p)) = (2+o(1))\lambda n^{2/3}$ a.a.s.

If we consider the dynamical model, Łuczak also proved that when $\lambda \to \infty$, the largest component in $G(n,p)$ will a.a.s. remain the largest until the end of the process, while when $\lambda \to -\infty$, the largest component will a.a.s. *not* remain the largest. When λ tends to a constant, the probability that the largest component remains the largest is bounded away from 0 and 1; the larger λ is, the closer the probability is to 1. For a detailed description of this phase of the process, see Janson and Spencer [26].

The process $G(n,p)$ obeys an interesting symmetry rule. Suppose that $p = \frac{c}{n}$ with $c > 1$, and let $d < 1$ be such that $de^{-d} = ce^{-c}$. Let C be the giant component in $G(n,p)$. The structure of $G(n,p) \setminus C$ is essentially that of $G(n',p')$, where n' is the number of vertices outside the giant component, and $p' = \frac{d}{n}$.

This symmetry rule has a parallel when $np \to 1$. Suppose that $np = 1 + \lambda n^{-1/3}$ with $\lambda \to \infty$, but $\lambda n^{-1/3} = o(1)$. At this point a giant component, C, has a.a.s. appeared in $G(n,p)$. Then the structure of $G(n,p) \setminus C$ is essentially similar to the structure of $G(n',p')$, where $n' = n - |C|$ and $n'p' = 1 - \lambda n^{-1/3}$.

The situation is very similar in the random graph model $G(n,m)$. In fact, this is the model for which Erdős and Rényi first described the phase transition. If $m = \frac{cn}{2}$ with $c < 1$, then the largest component in $G(n,m)$ a.a.s. has $O(\log n)$ vertices. If $m = \frac{cn}{2}$ with $c > 1$, then there is a.a.s. a unique component with $\Theta(n)$ vertices, and every other component has $O(\log n)$ vertices.

1.5 Random graphs with a given degree sequence

The random graph models $G(n,p)$ and $G(n,m)$ are by far the best understood, but many other ways of generating random graphs have been suggested. One of them is to choose a random graph with a given degree sequence, or as a special case, to choose a random regular graph. Newman, Strogatz and Watts [45] present several real-world graphs, which they demonstrate can be well approximated by this graph model.

A sequence $\mathbf{d} = (a_1, a_2, \ldots, a_n)$ of integers is called a *degree sequence* if $\sum_{i=1}^{n} a_i$ is even, and $0 \le a_i \le n-1$ for all $i = 1, \ldots, n$. We let $\Omega_{\mathbf{d}}$ be the set of all graphs on n vertices with degree sequence $\mathbf{d}$. Provided that $\Omega_{\mathbf{d}} \neq \emptyset$, we say that a random graph with degree sequence $\mathbf{d}$ is a graph chosen uniformly at random from $\Omega_{\mathbf{d}}$.

Since we are mostly interested in random graph models for which we can prove asymptotic results as n tends to infinity, we should define this random graph model for increasing n. We will mostly use the model and the terminology used by Molloy and Reed [44].

Let $A \subseteq \mathbb{N}$ be an infinite set of positive integers. An *asymptotic degree sequence* is a sequence of functions $\mathcal{D} = (d_0(n), d_1(n), d_2(n), \ldots)$, where $d_i : A \to \mathbb{N}_0$ for every $i \geq 0$, such that $d_i(n) = 0$ whenever $i \geq n$, and $\sum_{i \geq 0} d_i(n) = n$. If $\mathcal{D}$ is an asymptotic degree sequence and $n \in A$, we let $\mathcal{D}_n$ be the degree sequence $(a_1, a_2, \ldots, a_n)$, where $a_j \leq a_{j+1}$ for every $j = 1, \ldots, n-1$, and $\#\{j | a_j = i\} = d_i(n)$. Thus, $d_i(n)$ denotes the number of vertices of degree i in a graph of order n. The asymptotic degree sequence $\mathcal{D}$ is said to be *feasible* if $\Omega_{\mathcal{D}_n} \neq \emptyset$ for all $n \in A$.

Suppose that $\mathcal{D}$ is a feasible asymptotic degree sequence. If $n \in A$, a random graph $G_n(\mathcal{D})$ is a graph chosen uniformly at random from the set $\Omega_{\mathcal{D}_n}$. The graph $G_n(\mathcal{D})$ is called a *random graph with the given degree sequence* $\mathcal{D}_n$.

In order to be able to state interesting theorems about graphs in this random graph model, we need to impose some structure on $\mathcal{D}$. One way is to consider only random regular graphs: for some $r \geq 0$, $d_r(n) = n$ and $d_i(n) = 0$ for $i \neq r$, and A is restricted to the even numbers if r is odd. We will come back to this model in Section 1.5.3.

Another way, which allows for more general graphs, is to assume that the proportion of vertices of any given order is roughly the same for all $n \in A$. We will say that an asymptotic degree sequence is *smooth* if there are constants λ_i^* for $i \geq 0$ such that $\lambda_i(n) := d_i(n)/n \to \lambda_i^*$ as $n \to \infty$ for all $i \geq 0$. We will consider this model in Section 1.5.2.

1.5.1 The configuration model

It is difficult to study random graphs with a given degree sequence directly. Instead, it has become customary to take the route via random configurations. The configuration model was introduced by Bender and Canfield [6] and Bollobás [7], and later examined closer by Bollobás [11] and Wormald [58].

Given a degree sequence $\mathbf{d} = (a_1, \ldots, a_n)$, we define a configuration with degree sequence $\mathbf{d}$ in the following way. Let $V = \{v_1, \ldots, v_n\}$ be a set of vertices. Let L be a set consisting of a_i distinct copies of the vertex v_i for $i = 1, \ldots, n$. These copies are called *half-edges*. A *configuration* $\mathcal{C}$ consists of the set L, together with a perfect matching $\mathcal{P}$ of the half-edges in L. A *random configuration* $\mathcal{C}$ based on the set L is a configuration, in which the perfect matching $\mathcal{P}$ is chosen uniformly at random from the set of all perfect

matchings of $\mathcal{C}$.

A random perfect matching can be constructed greedily: at every step we take an arbitrary, unmatched half-edge, and match it with another half-edge chosen uniformly at random from the remaining half-edges. Using this procedure, every perfect matching has the same probability of being generated.

Given a configuration $\mathcal{C}$ on L, we define the underlying multigraph G^* of $\mathcal{C}$ to be the multigraph obtained by identifying all the copies of v_i with each other for $i = 1, \dots, n$. For an asymptotic degree sequence $\mathcal{D}$ we let $G_n^*(\mathcal{D})$ be the underlying multigraph of a random configuration $\mathcal{C}_n$ with degree sequence $\mathcal{D}_n$. Figure 1.1 shows a randomly generated configuration with degree sequence $(1, 1, 1, 2, 2, 2, 3, 3, 4, 5)$ and its underlying multigraph. The graph $G_n^*(\mathcal{D})$ is a random multigraph, but is not chosen uniformly at random from the set of multigraphs with degree sequence $\mathcal{D}_n$: the probability that $G_n^*(\mathcal{D}) = G$, where G is a multigraph with degree sequence $\mathcal{D}_n$ is proportional to $2^{-l} \prod_j j!^{-m_j}$, where l is the number of loops, and m_j is the number of multiedges with multiplicity j in G. (See for example Section 9.1 in [27].) This means, however, that $\mathbb{P}[G_n^*(\mathcal{D}) = G]$ is the same for any *simple* graph G. Thus, if we repeat the above procedure until we obtain a simple graph, we have generated a simple graph with degree sequence $\mathcal{D}_n$ uniformly at random. The configuration model can therefore be used to generate the graph $G_n(\mathcal{D})$.

If this procedure is to be used to generate random simple graphs, it is important that the probability that G_n^* is simple is not too small. If the probability that G_n^* is simple tends to 0 as n tends to infinity, the expected number of times the procedure must be repeated to obtain a simple graph increases with n. If we impose certain restrictions on $\mathcal{D}$, we can ensure that G_n^* is simple with probability bounded away from 0. This holds in particular when the maximum degree is bounded. In this case, if G_n^* has the property $\mathcal{P}$ a.a.s., then a simple random graph with degree sequence $\mathcal{D}$ also has $\mathcal{P}$ a.a.s.

1.5.2 The phase transition in random graphs with a given degree sequence

Molloy and Reed [44] showed that there is a phase transition for random graphs with given asymptotic degree sequence $\mathcal{D}$. We will assume that $\mathcal{D}$ is smooth — that is $\lambda_i(n) := d_i(n)/n \to \lambda_i^*$ for $i \geq 0$ — and that it is *sparse*, which means that $\sum_i d_i(n) = O(n)$. We require moreover that $\mathcal{D}$ should be "well-behaved" in a way which we will define precisely in Chapter 5, see page 82. In particular the maximum degree should not be too large. Molloy

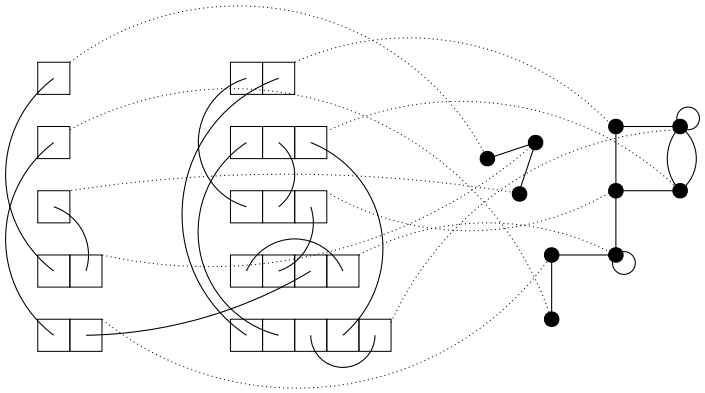

Figure 1.1: A configuration and its multigraph

and Reed defined the quantity

$$Q(\mathcal{D}) = \sum_{i \geq 1} i(i-2)\lambda_i^*,$$

and proved the following theorem about the phase transition in this graph model.

Theorem 1.3 (Molloy, Reed 1995)**.** *Let $\mathcal{D}$ be a well-behaved sparse asymptotic degree sequence for which there exists $\varepsilon > 0$ such that for all n and $i > n^{1/4-\varepsilon}$, $d_i(n) = 0$, and let $G = G_n(\mathcal{D})$. Then:*

(i) If $Q(\mathcal{D}) < 0$, and for some function $0 \leq \omega(n) \leq n^{1/8-\varepsilon}$, $d_i(n) = 0$ for all $i \geq \omega(n)$, then for some constant R dependent on $Q(\mathcal{D})$, G a.a.s. has no component with more than $R\omega(n)^2 \log n$ vertices, and a.a.s. has fewer than $2R\omega(n)^2 \log n$ cycles. Also, a.a.s. no component of G has more than one cycle.

(ii) If $Q(\mathcal{D}) > 0$, then there exist constants $\zeta_1, \zeta_2 > 0$ dependent on $\mathcal{D}$, such that G a.a.s. has a component with at least $\zeta_1 n$ vertices and $\zeta_2 n$ cycles. Furthermore, if $Q(\mathcal{D})$ is finite, then G a.a.s. has exactly one component of size greater than $C \log n$ for some constant C dependent on $\mathcal{D}$.

It may not be obvious why the quantity $Q(\mathcal{D})$ appears in Theorem 1.3. Suppose that we are given a randomly chosen vertex v in the graph and want to determine the order of the component it lies in. We can do this by exposing the component vertex by vertex. The vertex v has degree i with probability $\lambda_i(n)$. In this case there are i unexplored edges incident to v. If we follow one of these, we reach a new vertex (unless the edge is a loop); if this

vertex has degree j, the number of edges we can explore increases by $j-2$. We continue to explore the component, until at some point there are no longer any unexplored edges, in which case we have exposed the entire component. Whenever we follow an edge, the probability that the vertex we find at the other end has degree j is roughly $\frac{j\lambda_j(n)}{d}$, where d is the average degree. This holds as long as the number of explored vertices is small compared to the total number of vertices. Since the number of unexplored edges then increases with $j-2$, the expected increase in the number of unexplored edges is roughly $\frac{1}{d}\sum_j j(j-2)\lambda_j^*$. If this value is negative, we expect that the process will die out rather quickly; if it is positive, there is a chance that the number of unsaturated vertices will just continue to grow, so that a large component is generated.

In a subsequent paper, [43], Molloy and Reed also determined the order of the giant component: they found a function $\gamma(\mathcal{D})$ such that the giant component in G_n a.a.s. consists of $\gamma(\mathcal{D})n + o(n)$ vertices. Furthermore they proved that a duality principle holds for this graph model: the structure of the graph formed by removing the giant component from G_n is essentially the same as the structure of a random graph with asymptotic degree sequence $\mathcal{D}' = (d'_0(n), d'_1(n), \ldots)$, which can be calculated from $\mathcal{D}$.

The case that $Q(\mathcal{D}) = 0$, which is the *critical phase* of this random graph model, is not covered by Theorem 1.3; this will be the subject of Chapter 5. We will show that this graph model behaves roughly as $G(n,p)$ does in the critical phase.

1.5.3 Random regular graphs

A *random r-regular graph* is a graph chosen uniformly at random from the set of r-regular graphs. It is a special case of random graphs with a given degree sequence, where for all n, $d_r(n) = n$ and $d_i(n) = 0$ if $i \neq r$, and A is restricted to the even integers if r is odd. Because it is a natural model, it has mostly been studied in its own right, and not connected with the more general model described earlier in this section. The configuration model can be used to generate random regular graphs. If r is fixed, the graph produced by the configuration model is simple with probability bounded away from 0 when n grows. Hence, if the underlying multigraph of a random r-regular configuration has the property $\mathcal{P}$ a.a.s., a random simple r-regular graph also has the property $\mathcal{P}$ a.a.s.

Let $G_{\text{reg}}(n,r)$ be a random r-regular graph. For $r = 1$, $G_{\text{reg}}(n,r)$ is simply a perfect matching. For $r = 2$, $G(n,r)$ is a collection of cycles. It is fairly straightforward to show that $G_{\text{reg}}(n,2)$ is a Hamiltonian cycle with prob-

ability asymptotically equal to $\sqrt{\frac{\pi}{4n}}$; thus $G_{\text{reg}}(n,2)$ is a.a.s. disconnected. However, for $r \geq 3$, $G_{\text{reg}}(n,r)$ is a.a.s. r-connected. A problem which was open for a long time is whether $G_{\text{reg}}(n,r)$ a.a.s. contains a Hamiltonian cycle. This was settled in the affirmative by Robinson and Wormald, for $r = 3$ in [46] and for $r > 3$ in [47].

1.6 The minimum degree graph process

In Chapter 4 we will consider the minimum degree graph process, where the mechanism for adding edges guarantees that the minimum degree increases relatively quickly.

Let $\{G_{\min}(n,m)\}_{m\geq 0}$ be a Markov chain whose states are multigraphs on the set $\{1,2,\ldots,n\}$. The graph $G_{\min}(n,0)$ is the empty graph on n vertices, and for $m \geq 0$, $G_{\min}(n,m+1)$ is obtained from $G_{\min}(n,m)$ by first choosing a vertex of minimum degree in $G_{\min}(n,m)$ uniformly at random, and then connecting it by a new edge to another vertex chosen uniformly at random among the remaining vertices in $G_{\min}(n,m)$. Thus, at every step at least one vertex of minimum degree has its degree increased.

The process was originally introduced by Wormald [57] to illustrate the usage of the differential equation method, which we will come back to in Section 2.3. Kang et al. [30] later found the connectivity threshold for the process, and in Chapter 4 we will determine the point of the phase transition.

Let H_i for $i \geq 1$ be the random variables such that $G_{\min}(n, H_i - 1)$ has minimum degree less than i and $G_{\min}(n, H_i)$ has minimum degree at least i. Kang et al. [30] proved that there are constants $h_1 \approx 0.69$, $h_2 \approx 1.22$ and $h_3 \approx 1.73$ such that a.a.s. $H_1 = h_1 n + o(n)$, $H_2 = h_2 n + o(n)$ and $H_3 = h_3 n + o(n)$. (Exact formulas for h_1, h_2 and h_3 are given on page 49.) They moreover proved that if $t = \frac{m}{n}$, then if $t < h_2$, the graph is a.a.s. disconnected, while if $t > h_3$, the graph is a.a.s. connected. When $h_2 < t < h_3$, the probability that the graph is connected tends to a value bounded away from both 0 and 1 as n tends to infinity. In the case that the graph is not connected in this period, it a.a.s. consists of a giant component with $n - o(n)$ vertices, and one or more isolated cycles. In Section 4.1 we will find an exact expression for the limit probability that $G_{\min}(n, tn)$ is connected when $h_2 < t < h_3$.

In Chapter 4 we will prove that there is a phase transition in $G_{\min}(n,m)$, as there is in $G(n,m)$. We will prove that there is a constant $h_g \approx 0.86$ such that if $t < h_g$, then the graph a.a.s. consists of only small components, while if $t > h_g$, there is a giant component. In addition, we will show that the phase transition happens as a double jump: first the order of the largest component jumps from $O(\log n)$ to $\Theta(n^{2/3})$, and then from $\Theta(n^{2/3})$

to $\Theta(n)$. The behaviour of the phase transition is therefore similar to the phase transition in $G(n, m)$.

1.7 Other random graph models

A graph process similar to the minimum degree graph process is the so-called *min-min process*. In this process we also start with an empty graph, but at every step we choose *two* vertices of minimum degree, and add an edge between them. This is an attempt to define a graph process which produces random regular graphs, since after $rn/2$ edges have been added, the graph is necessarily r-regular, provided that rn is even. For $r = 1$, we get a perfect matching chosen uniformly at random. However, for $r = 2$, the resulting graph does not have the same distribution as the random regular graph $G_{\text{reg}}(n, 2)$. For $r \geq 3$, it is not known whether the min-min graph process has the same distribution as $G_{\text{reg}}(n, r)$. The min-min process is closely studied by Coja-Oghlan and Kang [14]. They find that as soon as the number of edges in the process is $(1+\varepsilon)n$, for any $\varepsilon > 0$, there is a.a.s. a giant component with more than $\frac{n}{2}$ vertices, and there is a positive probability that the graph is connected.

Another process which can be used to produce regular graphs is the d-process. In this process edges are added at random, subject to the restriction that the maximum degree should remain at most d. Every edge which is not already in the graph, and whose addition to the graph does not increase the degree of any vertex to larger than d has the same probability of being added. The process ends when no further edge can be added. Ruciński and Wormald [48] show that the final graph in this process is a.a.s. d-regular if nd is even, and has a.a.s. $n-1$ vertices of degree d and one vertex of degree $d-1$ if nd is odd. They also show in [49] that the final graph is a.a.s. connected when $d \geq 3$. For $d = 2$ this is not the case: Telcs, Wormald and Zhou [53] show that the graph is Hamiltonian, and thereby connected, with probability $\Theta(n^{-1/2})$, although with a different constant than in the uniformly random 2-regular graph.

In recent years there has been an interest in studying real-world networks and modelling these with random graph processes. This was sparked by a paper by Watts and Strogatz [54] from 1998. They consider some real-world networks — the neural network of a roundworm, the power grid of the Western United States, and the collaboration graph for movie actors — and show that these graphs share some characteristic properties, namely high clustering along with small diameter, in spite of the graphs being sparse. They call these "small-world" networks, in analogy with the "small-world"

phenomenon, popularised by the notion of "six degrees of separation". The random graphs which we have presented so far in this chapter typically have small diameters, but lack the clustering property, whereas graphs which exhibit more regularity, such as lattices, are generally clustering, but have large diameters. Thus, those models are inadequate for the purpose of studying these real-world networks, and researchers have attempted to find random graph models which typically produce graphs that share these traits, hoping to achieve mathematical models which are useful in the study of real-world networks.

Another real-world network that has been given much attention, is the "web graph", in which the vertices are pages on the World Wide Web, and links between pages are represented by edges. Albert, Barabási and Jeong [1] showed that the web graph has relatively small diameter, while Faloutsos, Faloutsos and Faloutsos [19] found that its degree sequence obeys a power law: the number of vertices of degree d is proportional to d^a for some constant a. Graphs satisfying such a law are often called *scale-free*, and several other real-world networks have also been shown to closely follow such power laws.

Subsequently it has been attempted to design random graph models which produce scale-free graphs having the "small-world" properties mentioned above. A survey of such scale-free random graphs is found in [12]. One such scale-free graph process was proposed by Barabási and Albert [5]. In this process one starts with some small arbitrary graph, and adds vertices one by one to the graph. Every vertex added is connected to some of existing vertices, in such a way that the probability that a vertex receives a new edge is proportional to its degree.

1.8 Random digraphs

Finally we will consider another random structure, namely random digraphs. In the binomial model for random digraphs, $D(n,p)$, a random digraph on n vertices is chosen such that for every ordered pair (v,w) of vertices there is an arc from v to w with probability p. All pairs are considered independently of one another. In the context of digraphs we will use the term *component* to mean a strongly connected component.

Both Karp [31] and Łuczak [38] considered the phase transition, and proved that it occurs when $p = 1/n$. That is, let $p = c/n$, where c is a constant. If $c < 1$, then a.a.s. all the components in $D(n,p)$ are single vertices or cycles of length at most $\omega(n)$, for any function $\omega(n) \to \infty$. If $c > 1$, then there is a.a.s. a giant component with $\Theta(n)$ vertices. In fact, the

order of the giant component is a.a.s. $(d^2 + o(1))n$, where d is the constant appearing in Theorem 1.1(iii). Thus the probability that a specified vertex belongs to the giant component in $D(n,p)$ is the square of the probability that an arbitrary vertex belongs to the giant component in $G(n,p)$.

In Chapter 6 we will consider the critical phase in the evolution of random digraphs, namely when $c = 1$. We will prove that using the same parametrisation as in $G(n,p)$, namely (1.1), the random digraph exhibits the same behaviour as $G(n,p)$: if $\lambda \to -\infty$, then a.a.s. all components in $D(n,p)$ are relatively small cycles or single vertices, while if $\lambda \to \infty$, then there is a.a.s. exactly one component which is neither a cycle nor a vertex, which is much bigger than every other component.

1.9 Summary of the main results

The main results in this dissertation regard the phase transitions in three different random graph models: the minimum degree graph process, random graphs with a given degree sequence and random digraphs. In the first of these random graph models we will locate the point of the phase transition and show that the graph process undergoes a *double jump*, similar to $G(n,p)$. In the two other random graphs, for which the phase transitions have already been located ([44], [31], [38]), we will consider how the random graphs behave very close to the point of the phase transition. The three random graph models will be treated in Chapters 4, 5 and 6 respectively. Below we state the main theorem in each of these chapters.

The minimum degree random graph process $G_{\min}(n,m)$ was introduced in Section 1.6, and will mainly be treated in Chapter 4, although Chapter 3 also considers some of its aspects. We will prove that the phase transition occurs when roughly $h_g n$ edges have been added, for a constant h_g.

Theorem 4.2. *Let*

$$h_g = \log \frac{16\log 2 - 2}{3\log 2 - 1 + \log 2 \cdot \sqrt{27 - 16\log 2}} \approx 0.8607.$$

(i) If $t < h_g$, then a.a.s. every component in $G_{\min}(n,tn)$ has $O(\log n)$ vertices.

(ii) If $t = h_g$ and $\omega(n) \to \infty$, then $G_{\min}(n,tn)$ a.a.s. contains no component of order greater than $n^{2/3}\omega(n)$, and at least one component of order greater than $n^{2/3}/\omega(n)$.

(iii) If $t > h_g$, then a.a.s. the largest component in $G_{\min}(n, tn)$ has $s(t)n + o(n)$ vertices, where $s(t)$ is a function depending only on t, and every other component, if any, has $O(\log n)$ vertices.

Random graphs with a given degree sequence $G_n(\mathcal{D})$ were presented in Section 1.5, and the point of the phase transition was already found by Molloy and Reed [44], see Theorem 1.3. In Chapter 5 we will consider more closely the critical point itself. In order to state the main theorem of Chapter 5, we will need to introduce some more notation. Recall that $\lambda_i(n)$ is the proportion of vertices of degree i in a graph with n vertices. Let

$$Q_n(z) = \sum_{i \geq 1} i(i-2)\lambda_i(n) z^i,$$

and let τ_n be the largest value such that

$$Q_n(\tau_n) = 0.$$

We will prove the following theorem.

Theorem 5.2. *Assume that $\mathcal{D}$ is a well-behaved asymptotic degree sequence, such that for some $\varepsilon > 0$, $d_i(n) = 0$ whenever $i > n^{1/4-\varepsilon}$. Furthermore assume that $\lim_{n\to\infty} \tau_n = 1$ and $\lambda_1^* > 0$. Let*

$$\delta_n = 1 - \tau_n.$$

(i) If $\delta_n n^{1/3} \to -\infty$, then a.a.s. all components in $G_n(\mathcal{D})$ have $o(n^{2/3})$ vertices.

(ii) There is a constant c_1 such that if $\delta_n n^{1/3} \geq c_1 \log n$ then a.a.s. $G_n(\mathcal{D})$ has a single component with $\gg n^{2/3}$ vertices, while all other components have $o(n^{2/3})$ vertices.

Finally, in Chapter 6, we will consider the random digraph $D(n,p)$, which we introduced in Section 1.8. It was proved by Karp [31] and Łuczak [38] that the phase transition in $D(n,p)$ happens when $p = \frac{1}{n}$, which is the same as for $G(n,p)$. The next natural question is whether one can prove an analogue of Theorem 1.2 for random digraphs when $np \to 1$. We will prove the following theorem, which shows that the situation for random digraph has similarities to the random undirected graph.

Theorem 6.1. *Let $np = 1 + \varepsilon$, such that $\varepsilon = \varepsilon(n) \to 0$, and let $\omega(n) \to \infty$.*

(i) If $\varepsilon^3 n \to -\infty$, then a.a.s. every component in $D(n,p)$ is either a vertex or a cycle of length at most $\omega(n)/|\varepsilon|$.

(ii) If $\varepsilon^3 n \to \infty$, then a.a.s. $D(n,p)$ contains a unique complex component, which has order $(4 + o(1))\varepsilon^2 n$, while every other component is either a vertex or a cycle of length at most $\omega(n)/\varepsilon$.

Chapter 2

Tools

In this chapter we will present some important results from different areas of mathematics, which will prove useful in later chapters.

In Section 2.1 we present the concept of generating functions, and a theorem about the asymptotic growth of the coefficients of certain generating functions, which can be derived using singularity analysis of complex functions.

In Section 2.2 we present some basic results from the theory of branching processes. We will then sketch a proof of parts (i) and (iii) of Theorem 1.1 using a branching process. In Chapters 4 and 5 we will use branching processes to study the phase transitions in the minimum degree graph process from Section 1.6 and in random graphs with a given degree sequence from Section 1.5, respectively. However, the branching process argument is easier and more transparent in the case of $G(n, p)$, and it is therefore useful to consider this case first, so that it can serve as a model for the discussion of the other processes.

In Section 2.3 we present a method using differential equations to study discrete random variables defined on graph processes, and Section 2.4 contains a martingale central limit theorem, which we will use in Chapter 3 to show asymptotic joint normality of a sequence of random variables.

2.1 Generating functions

Generating functions will be used throughout this dissertation, often because their usage significantly simplifies calculations, but sometimes we will also use properties of the generating functions themselves. We therefore include a short introduction of generating functions and present a theorem which provides a connection between the asymptotic behaviour of a sequence and

the singularities of the corresponding generating function considered as a function in the complex plane. A nice exposition of generating functions has been written by Wilf [55]. A more comprehensive book, which especially focuses on the analytic aspects of generating functions, is being written by Flajolet and Sedgewick [20].

Given a sequence $(a_0, a_1, a_2, \ldots)$ of real numbers, the *generating function* of the sequence is the formal power series

$$f(z) = \sum_{k \geq 0} a_k z^k.$$

Given a generating function $f(z)$, we use the notation $\left[z^k\right] f(z)$ to denote the coefficient of z^k in $f(z)$, so in this example $\left[z^k\right] f(z) = a_k$.

We will often work with *probability generating functions*. Let X be a random variable which takes nonnegative integers as values, and let $p_k = \mathbb{P}[X = k]$ for $k \geq 0$. Then the probability generating function corresponding to X is

$$p(z) = \sum_{k \geq 0} p_k z^k.$$

Usually we have $p(1) = 1$; in some cases, however, we allow the event $X = \infty$ to have nonzero probability, and in this case $p(1) = \mathbb{P}[X < \infty] < 1$. Several properties of X are easy to derive from the probability generating functions. Assume that $p(1) = 1$. Then the expected value of X is

$$\mathbb{E}[X] = \sum_{k \geq 0} k p_k = p'(1),$$

and the variance is

$$\mathrm{Var}[X] = p''(1) + p'(1) - p'(1)^2.$$

Given the generating function $f(z)$ of a sequence $\{a_0, a_1, \ldots\}$, it can be used to determine the asymptotic growth of the sequence. Let ρ be the radius of convergence of $f(z)$: if z is a point in the complex plane, then $\sum_k a_k z^k$ converges whenever $|z| < \rho$, and diverges whenever $|z| > \rho$. Then $a_n = \rho^{-n}\theta(n)$, where $\theta(n)$ is a subexponential factor. Sometimes the notation $a_n \bowtie \rho^{-n}$ is used to denote this.

The formal power series can then be viewed as a function in the complex plane, which is analytic in a disc of radius ρ around the origin. We know from complex analysis that there is a unique analytic continuation of the function to the entire complex plane, and we also know that this function must have a singularity on the circle $|z| = \rho$. Furthermore, Pringsheim's

theorem (see [20]) asserts that if a_n is nonnegative for all n, then $z = \rho$ is a singularity of the complex function $f(z)$. If we are given the function $f(z)$, we can determine the asymptotic growth of the coefficients of its Taylor series, by finding the singularity of $f(z)$ which is closest to the origin. This singularity is referred to as the *dominant singularity* of $f(z)$.

Sometimes one also wants to find the subexponential function $\theta(n)$. For many functions this can be derived from the behaviour of $f(z)$ near its dominant singularity. We will only consider the following case, where the function $y(z)$ is a function defined implicitly by $y(z) = z\phi(y(z))$, and $\phi(z)$ is a known function. The following theorem is a simplified version of Theorem VI.6 in [20]. A function $\phi(w)$ is *periodic* if, for some $d \geq 2$ and function ψ, we have $\phi(w) = \psi(w^d)$.

Theorem 2.1. *Let $\phi(w)$ be a function analytic at 0, having nonnegative Taylor coefficients with $\phi(0) \neq 0$, such that there exists a positive solution τ to the characteristic equation*

$$\phi(\tau) - \tau\phi'(\tau) = 0$$

strictly within the disc of convergence of ϕ. Let $y(z)$ be the solution analytic at the origin of $y(z) = z\phi(y(z))$. Then $y(z)$ has a dominant singularity at $z = \rho$, where $\rho = \frac{\tau}{\phi(\tau)}$. If, furthermore, $\phi''(\tau) \neq 0$ and $\phi(w)$ is aperiodic, then the coefficients of $y(z)$ satisfy

$$\left[z^i\right] y(z) \sim c\rho^{-i} i^{-\frac{3}{2}} \left(1 + O(i^{-1})\right), \tag{2.1}$$

where $c = \sqrt{\frac{\phi(\tau)}{2\pi\phi''(\tau)}}$.

In Chapters 4 and 5 we will use Theorem 2.1, combined with an analysis of certain generating functions close to $z = 1$. For this we will need the following lemma.

Lemma 2.2. *Let $f_n(x) = \sum_{i\geq 0} \alpha_{ni} x^i$, where α_{ni} are real numbers for $n \geq 1$ and $i \geq 0$. Assume that $\alpha_{ni} \to \alpha_i^*$ as $n \to \infty$ for constants α_i^* for all $i \geq 0$. Let $r(n) = \max(\{i : \alpha_{ni} \neq 0\} \cup \{0\})$. Let $\{a_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$ be sequences of real numbers such that $a_n \to 1$ and $b_n \to 0$ as $n \to \infty$. Assume that $r(n)b_n = o(1)$. Then, as $n \to \infty$,*

$$f_n(a_n + b_n) = f_n(a_n) + b_n f_n'(a_n) + \frac{1}{2} b_n^2 f_n''(a_n) + O(b_n^3). \tag{2.2}$$

Proof.

$$\begin{aligned} f_n(a_n + b_n) &= \sum_{i\geq 0} \alpha_{ni}(a_n + b_n)^i \\ &= \sum_{i\geq 0} \left(\alpha_{ni} \sum_{j=0}^{i} \binom{i}{j} a_n^{i-j} b_n^j \right) \\ &= \sum_{i\geq 0} \alpha_{ni} \left(a_n^i + i a_n^{i-1} b_n + \frac{1}{2} i(i-1) a_n^{i-2} b_n^2 + O(b_n^3) \right) \\ &= f_n(a_n) + b_n f_n'(a_n) + \frac{1}{2} b_n^2 f_n''(a_n) + O(b_n^3). \end{aligned}$$

□

2.2 Branching processes

A branching process is a random process which examines some population of organisms, or particles. The branching processes we will consider proceed in discrete steps, which we call *generations*. Every individual particle in the gth generation has a random number of children in the $g+1$st generation. When we use branching processes in a graph context, we will generally call the particles "vertices". Two comprehensive books about branching processes are Harris [22] and Athreya and Ney [3].

A central theme is whether a branching process dies out or continues growing forever. The *extinction probability* of a branching process is the probability that it dies out after a finite number of steps. Branching processes can be classified in three categories: subcritical, critical and supercritical. In *subcritical* and *critical* branching processes the extinction probability is 1, while *supercritical* branching processes are those where the extinction probability is strictly smaller than 1. The difference between subcritical and critical branching processes, is that the expected number of generations before the branching process dies out is finite in the subcritical case and infinite in the critical case.

2.2.1 The Galton-Watson process

In the Galton-Watson process every particle has a number of children which is distributed according to a nonnegative, integral random variable X, which is the same for every particle in every generation. The process starts with one particle in generation 0. We will assume that X has finite expectation.

We let Z_n be the random variables denoting the number of particles in generation n. Thus $Z_0 = 1$ and for $n \geq 0$,

$$Z_{n+1} = \sum_{i=1}^{Z_n} X_{ni},$$

where X_{ni} has the same distribution as X for $n = 0, 1, \ldots$ and $1 \leq i \leq Z_n$. Let $p(z) = \sum_{k \geq 0} \mathbb{P}[X = k] z^k$ be the probability generating function of X.

According to the following fundamental theorem, Theorem I.5.1 in [3], whether the Galton-Watson process is supercritical or not depends only on the expected value of X, and the extinction probability can be calculated from the generating function $p(z)$.

Theorem 2.3. *The extinction probability of the $\{Z_n\}$ process is the smallest nonnegative root of the equation*

$$p(z) = z.$$

It is 1 if $\mathbb{E}[X] \leq 1$ and smaller than 1 if $\mathbb{E}[X] > 1$.

2.2.2 Branching process argument for $G(n, p)$

In this section we will present a branching process argument which can be used to prove part (i) and (iii) of Theorem 1.1. The essence of the argument is well-known, see for example [2]. In our presentation we will use Theorem 2.1 to derive part (i) and Theorem 2.3 to show part (iii).

Assume that $p = \frac{c}{n}$, where c is a positive constant. We take an arbitrary vertex v in $G = G(n, p)$ and want to determine the order of the component containing v. We expose this component by using a sequential search in the following way. We say that vertices can be *live*, *dead* or *neutral*. In the beginning v is the only live vertex, while every other is neutral. At every step we choose a live vertex w, and for every neutral vertex x we check whether the edge wx is contained in the graph. If it is, then we mark x as live, and when all neutral vertices have been checked in this way, we mark w as dead. If there at some point is no live vertex left, we have exposed the entire component containing v.

Every time we ask whether the graph contains a particular edge, we get an affirmative answer with probability p. By the above procedure we are guaranteed that we will never ask for the same edge twice. Hence every edge we check is in the graph with probability p independently of every other edge.

Let Y_k be the random variable denoting the number of exposed vertices, (that is the number of vertices which are either live or dead) after step k in

this process. We have $Y_0 = 1$, since before the first step, the only exposed vertex is v. We let Z_k be the number of vertices exposed in step k. Then $Y_k = Y_{k-1} + Z_k$.

The number of vertices found in the kth step is distributed binomially, as $\mathrm{Bin}(n - Y_k, p)$. Provided that Y_k is small compared to n, the expected number of vertices exposed in the kth step is $(n - Y_k)p \sim np = c$. As n tends to infinity, the binomial distribution converges to the Poisson distribution with mean c. Thus, the probability that the number of vertices exposed in the kth step is i, tends to $\frac{e^{-c}c^i}{i!}$ as n tends to infinity, as long as Y_k is not too big. The process of exposing the component which contains the vertex v is thus very similar to a Galton-Watson process where the number of children of a particle is a random variable with Poisson distribution with mean c. There are, however, certain differences. In the branching process the distribution for the number of children remains the same throughout the process, while in the exposure process it changes as there are fewer and fewer neutral vertices left. Moreover, the branching process can potentially go on forever, while in the exposure process there are only n vertices available. Nevertheless, the approximation is so good that one can prove theorems regarding the component structure of $G(n, p)$ using information about the branching process.

Let us now consider the Galton-Watson process where every vertex has a number of children which is distributed according to the Poisson distribution with mean c. The probability generating function is

$$p(z) = \sum_{k \geq 0} \frac{e^{-c}c^k}{k!} z^k = e^{c(z-1)}.$$

Let q_k be the probability that the branching process dies out after producing precisely k vertices, including the vertex we started with, and let $q(z) = \sum_{k \geq 0} q_k z^k$ be the corresponding probability generating function. The process is self-similar in the sense that if we take an arbitrary vertex v at any point in the process, then the random variable denoting the total number of descendants of v, including v itself, has the same distribution as the random variable denoting the total number of vertices in the entire process. Thus, if v is a vertex with k children, then the probability generating function for the number of descendants of v, including v, is $zq(z)^k$. It follows that

$$q(z) = \sum_{k \geq 0} p_k z q(z)^k = zp(q(z)).$$

Thus, $q(z)$ is implicitly defined in a form we recognise from Theorem 2.1, with $\phi(z) = p(z)$ and $y(z) = q(z)$. We have

$$p(z) - zp'(z) = (1 - cz)e^{cz-1},$$

so $\tau = \frac{1}{c}$ satisfies the equation $p(\tau) - \tau p'(\tau) = 0$. The dominant singularity of $q(z)$ is then $\rho = \frac{\tau}{p(\tau)} = \frac{1}{c}e^{c-1}$, according to Theorem 2.1. Hence $q_k \bowtie (ce^{1-c})^k$. If one wants to be more exact, Theorem 2.1 moreover implies that $q_k \asymp (ce^{1-c})^k k^{-3/2}$. Let us now calculate the probability that the branching process produces more than $a \log n$ vertices, where $a = \frac{1}{c-1-\log c}$. By Theorem 2.3, if $c < 1$, the branching process dies out after a finite number of steps with probability 1. Hence, in this case

$$\begin{aligned}
\mathbb{P}[Y \geq a \log n] &= \sum_{k \geq a \log n} q_k \asymp \int_{a \log n}^{\infty} (ce^{1-c})^x x^{-3/2} dx \\
&= \int_{a \log n}^{\infty} e^{x(1-c+\log c)} x^{-3/2} dx \\
&\sim \left[-2e^{x(1-c+\log c)} x^{-1/2} \right]_{x=a \log n}^{\infty} \\
&= 2n^{a(1-c+\log c)} (a \log n)^{-1/2} = o(n^{-1}).
\end{aligned}$$

Thus, for every vertex v, the probability that v is in a component of order greater than $a \log n$ is $o(n^{-1})$, so the probability that there is any such component in $G(n,p)$ is $n \cdot o(n^{-1}) = o(1)$.

When $c > 1$, there is a positive probability that the branching process continues forever. According to Theorem 2.3, the extinction probability of the branching process is the smallest nonnegative root of the equation $p(z) = z$. The solution to this equation is $1 - d$, where d is the constant appearing in part (iii) of Theorem 1.1. One can show that in the case that the branching process does die out, the probability that it generates more than $O(\log n)$ vertices before dying out is $o(n^{-1})$.

Let v be a vertex in $G(n,p)$, and suppose we expose the component containing v as explained above. Then, with probability $1 - d + o(1)$, the process dies out after at most $O(\log n)$ vertices have been exposed, while with probability $d + o(1)$, the process finds more than $O(\log n)$ vertices. To prove that there is a giant component, which we will not do here, one has to show that the vertices, for which this process exposes many vertices, all join together in a single component.

2.2.3 Multitype branching processes

A multitype branching process is a generalisation of the Galton-Watson process. We have a finite number of types of particles, and the number of children a particle has, is a random variable which depends only on the type of the particle.

Suppose that there are k different types. For $i = 1, \ldots, k$, we let $p^{(i)}(j_1, \ldots, j_k)$ be the probability that a particle of type i produces $j_{i'}$ particles

of type i' for $i' = 1, \ldots, k$. Let

$$f^{(i)}(z_1, \ldots, z_k) = \sum_{j_1, \ldots, j_k \geq 0} p^{(i)}(j_1, \ldots, j_k) z_1^{j_1} \cdots z_k^{j_k}$$

be the k-variate probability generating function associated with particles of type i for $i = 1, \ldots, k$.

As in the case of the Galton-Watson process, a central question is to determine whether a branching process dies out with probability equal to 1, or less than 1. To determine this, we introduce the *transition matrix* of a multitype branching process. The transition matrix A is a $k \times k$-matrix where the entry in the ith row and jth column equals the expected number of children of type j which are produced by a single particle of type i in one generation.

Suppose that there are v_i particles of type i for $i = 1, \ldots, k$ in some generation. We express this as a column vector $\mathbf{v} = [v_1, \ldots, v_k]'$. Then the expected number of particles of the different types m generations later is given by the vector $A^m \mathbf{v}$. It is therefore not surprising that the destiny of the branching process is related to the value of the largest eigenvalue of A. This is stated in the following theorem, which is Theorem V.3.2 in [3]. (The matrix A is *strictly positive* if, for some n, all the entries in A^n are positive.)

Theorem 2.4. *Assume that A is strictly positive, and let λ_1 be the largest eigenvalue of A. If $\lambda_1 \leq 1$, then the branching process dies out after a finite number of steps with probability 1. If $\lambda_1 > 1$, then there is a positive probability that the branching process continues forever.*

In the latter case, the set of equations

$$\begin{aligned} y_1 &= f^{(1)}(y_1, \ldots, y_k) \\ &\vdots \\ y_k &= f^{(k)}(y_1, \ldots, y_k) \end{aligned}$$

has a unique solution satisfying $0 \leq y_i < 1$ for $i = 1, \ldots, k$. Then y_i is the probability that a branching process starting with a single particle of type i dies out after a finite number of steps.

2.3 Differential equations

In this section we consider random Markovian processes which proceed in discrete steps. In our setting we have a sequence of random processes indexed by n, and we want to study the asymptotic behaviour of random variables

defined on these processes, as n tends to infinity. The technique we will present involves a passage from the discrete to the continuous point of view. We introduce a time variable t, in such a way that every time interval of fixed length corresponds to a number of steps in the discrete process which is linear in n. Many random variables defined on the process can then be scaled in such a way that the expectations of the scaled variables converge to some fixed continuous functions, which emerge as the solutions of certain differential equations.

This technique was first used by Karp and Sipser [32] to analyse a random greedy matching algorithm. It was later tailored for random graph processes and popularised by Wormald [56]. Wormald also wrote a survey [57] of the differential equation method and some of its applications for random graphs.

In the context of graphs, we have a sequence of random graph processes indexed by the number of vertices, n, and we have certain random variables defined on the graphs. The graph processes proceed by having one edge added at every step, and the idea is to express the expected change in the random variables in terms of the values of the random variables after the previous step. From these expressions we derive differential equations, the solutions of which are close to the expected value of the random variables divided by n. Using a martingale argument one can show that if certain requirements are imposed on the random variables and the differential equations, the random variables are sharply concentrated around their expectations.

The following lemma is a version of Theorem 5.1 in [57], suitable for our purposes later on. In some respects it is slightly weaker than the theorem as it is presented in [57], but in one important respect it is stronger, namely in that the number of random variables, k_0, is allowed to increase with n. That the lemma holds in this case is stated in the note following Theorem 5.1 in [57].

Theorem 2.5. *Let $\{G_m\}_{m\geq 0}$ be a random graph process, whose stages are graphs on n vertices. For $k = 1, 2, \ldots, k_0$ with $k_0 = O(\log n)$, let $X_k(T)$ be a random variable defined on the process up to time T, $\{G_m\}_{m=0}^{T}$, for each $T \geq 1$. Suppose also that $|X_k(T)(G_m)| < Cn$ for some constant C, and that for some functions $m = m(n)$, and $f_k : \mathbb{R}^{k_0+1} \to \mathbb{R}$ the following hold:*

1. *there is a constant C' such that*

$$|X_k(T+1) - X_k(T)| \leq C'$$

for all $T < m$ and $k = 1, 2, , \ldots, k_0$,

2. *for* $k = 1, 2, \ldots, k_0$,

$$\mathbb{E}\left[X_k(T+1) - X_k(T) \mid G_T\right] = f_k(T/n, X_1(T)/n, \ldots, X_{k_0}(T)/n) + o(1)$$

uniformly over all $T < m$,

3. *for each* $k = 1, 2, \ldots, k_0$, *the function* f_k *is continuous and satisfies a Lipschitz condition on* D, *where* D *is some bounded connected open set containing the intersection of* $\{(s, z_1, \ldots, z_{k_0}) : s \geq 0\}$ *with some neighbourhood of* $\{(0, z_1, \ldots, z_{k_0}) : \mathbb{P}\left[X_k(0) = z_k n, 1 \leq k \leq k_0\right] \neq 0\}$.

Then,

(a) *for* $(T, \hat{z}_1, \ldots, \hat{z}_{k_0}) \in D$, *the system of differential equations*

$$\frac{dz_k}{ds} = f_k(s, z_1, \ldots, z_{k_0}), \qquad k = 1, 2, \ldots, k_0,$$

has a unique solution in D *for* $z_k : \mathbb{R} \to \mathbb{R}$ *passing through*

$$z_k(0) = \hat{z}_k, \quad k = 1, 2, \ldots, k_0,$$

and which extends to points arbitrarily close to the boundary of D.

(b) *for each* $k = 1, 2, \ldots, k_0$, *a.a.s.*

$$X_k(T) = n z_k(T/n) + o(n)$$

uniformly for $0 \leq T \leq \min\{\sigma n, m\}$, *where* $z_k(T)$ *is the solution in (a) with* $\hat{z}_k = X_k(0)/n$, *and* $\sigma = \sigma(n)$ *is the supremum of those* s *to which the solution can be extended.*

To illustrate how the theorem can be used, we use it to calculate the degree distribution of the minimum degree graph process from Section 1.6, until the minimum degree becomes 2. We refer to Section 1.6 for the definition of the process.

If $G_{\min}(n, m)$ contains an isolated vertex, the first vertex chosen in the next step has degree 0, and will therefore have its degree increased to 1. The second vertex chosen is chosen uniformly at random from the remaining vertices, and has therefore degree k with probability $\frac{X_k(n,m)-\delta_{0k}}{n-1}$. Thus, for $k \geq 0$,

$$\begin{aligned}\mathbb{E}\left[X_k(n, m+1) - X_k(n, m) | G_{\min}(n, m)\right] &= -\delta_{0k} + \delta_{1k} - \frac{X_k(n,m)}{n} \\ &\quad + \frac{X_{k-1}(n,m)}{n} + o(1),\end{aligned}$$

where $X_{-1}(n,m)$ is defined to be 0 for all $n \geq 1$ and $m \geq 0$. This equation has the same form as the equation in point 2 of Theorem 2.5, with $f_k(t, z_1, \ldots, z_{k_0}) = -\delta_{0k} + \delta_{1k} - z_k + z_{k-1}$. In order to show that the theorem can be used, we also have to show that points 1 and 3 are satisfied. For all k we have $|X_k(T+1) - X_k(T)| \leq 2$, so point 1 is satisfied. Since the functions f_k are all linear, they also satisfy a Lipschitz condition in a suitable set D.

Hence the system of differential equations

$$\frac{dz_k}{ds} = f_k(s, z_1, \ldots, z_{k_0})$$

has a unique solution in D. Let $(\alpha_0(t), \alpha_1(t), \ldots, \alpha_{k_0}(t))$ be the solution. Theorem 2.5 says that the random variables $X_k(n,m)$ are concentrated around $n\alpha_k(t)$, that is a.a.s. $X_k(n,m) = n\alpha_k(t) + o(n)$.

Thus, we only have to calculate the functions $\alpha_k(t)$. In order to do this, we define the generating function

$$A(z,t) = \sum_{k \geq 0} \alpha_k(t) z^k.$$

We know that the $\alpha_k(t)$'s satisfy the differential equation

$$\frac{d}{dt}\alpha_k(t) = -\delta_{0k} + \delta_{1k} - \alpha_k(t) + \alpha_{k-1}(t)$$

for $k \geq 0$, where we define $\alpha_{-1}(t) = 0$ for all $t \geq 0$. Multiplying this equation by z^k and summing over k, we obtain

$$\begin{aligned} \frac{\partial}{\partial t}A(z,t) = \sum_{k \geq 0} \frac{d}{dt}\alpha_k(t) z^k &= \sum_{k \geq 0}(-\delta_{0k} + \delta_{1k} - \alpha_k(t) + \alpha_{k-1}(t)) z^k \\ &= -1 + z - A(z,t) + zA(z,t) \\ &= (z-1)(A(z,t)+1). \end{aligned}$$

When $t = 0$, all vertices have degree 0, so $A(z,0) = 1$. Solving the differential equation, together with this boundary condition, we get

$$A(z,t) = 2e^{t(z-1)} - 1.$$

This holds as long as there are isolated vertices in the graph; that is, until $t = \log 2$. When $t = \log 2$, $A(z, \log 2) = 2^z - 1$, and the minimum degree becomes 1.

When the minimum degree is 1, the $X_k(n,m)$'s satisfy the equation

$$\begin{aligned}\mathbb{E}\left[X_k(n,m+1) - X_k(n,m) | G_{\min}(n,m)\right] &= -\delta_{1k} + \delta_{2k} - \frac{X_k(n,m)}{n} \\ &\quad + \frac{X_{k-1}(n,m)}{n} + o(1).\end{aligned}$$

Hence, for $h_1 < t < h_2$,

$$\frac{d}{dt}\alpha_k(t) = -\delta_{1k} + \delta_{2k} - \alpha_k(t) + \alpha_{k-1}(t).$$

Again multiplying by z^k and summing over k, we obtain

$$\frac{\partial}{\partial t}A(z,t) = (z-1)(A(z,t)+z).$$

If we solve this differential equation and use the boundary condition that $A(z, \log 2) = 2^z - 1$, we get $A(z,t) = \left(\frac{e^t}{2}\right)^{z-1}(z - 1 + 2^z) - z$. Thus,

$$A(z,t) = \begin{cases} 2e^{t(z-1)} - 1 & \text{if } t \le h_1, \\ \left(\frac{e^t}{2}\right)^{z-1}(z-1+2^z) - z & \text{if } h_1 < t \le h_2, \end{cases}$$

and from this we can deduce that

$$\alpha_k(t) = \begin{cases} \frac{2t^k}{k!e^t} - \delta_{k0} & \text{if } h_1 \le t, \\ \frac{2}{k!e^t}\left(t^k + (t-\log 2)^{k-1}(k + \log 2 - t)\right) - \delta_{k1} & \text{if } h_1 < t \le h_2, \end{cases}$$

where we write $0^0 = 1$.

2.4 Martingale central limit theorem

A discrete martingale is a sequence of random variables $X_0, X_1, \ldots$ such that for $i \ge 0$, $\mathbb{E}\left[X_{i+1} | X_0, \ldots, X_i\right] = X_i$. Martingales are often used for proving large deviation inequalities. In this section we will present a martingale central limit theorem, which is taken from Janson [24], who in turn based it on a theorem in Jacod and Shiryaev [23].

The martingales used in [24] have index set $[0, \infty)$ rather than $\{0, 1, \ldots\}$, as in our definition above. We will only use discrete martingales, and defining continuous martingales requires much terminology to be introduced, so we will refrain from defining them; however, we will state the theorem as it is stated in [24], as it is straightforward to apply in a discrete setting.

Let X and Y be two real-valued martingales of finite variation, defined on $[0,\infty)$. Let $\Delta X(s) = X(s) - X(s^-)$ and $\Delta Y(s) = Y(s) - Y(s^-)$. We then define

$$[X,Y]_t = \sum_{0<s\le t} \Delta X(s)\Delta Y(s).$$

The *quadratic variation* of the martingale X is $[X,X]_t$. If $X = (X_i)_{i=1}^m$ and $Y = (Y_j)_{j=1}^n$ are two vector-valued martingales, we define $[X,Y]_t$ to be the $m \times n$ matrix $([X_i, Y_j]_t)_{i,j}$.

Theorem 2.6 (Proposition 9.1(i) in [24])**.** *Assume that for each n, $M_n(x) = (M_{ni}(x))_{i=1}^q$ is a real q-dimensional martingale on $[0,\infty)$ with $M_n(0) = 0$, and that $\Sigma(x)$, $x \ge 0$, is a (nonrandom) continuous matrix-valued function such that for every fixed $x \ge 0$,*

$$[M_n, M_n]_x \xrightarrow{p} \Sigma(x) \text{ as } n \to \infty$$

and

$$\sup_n \mathbb{E}|M_n(x)|^2 < \infty.$$

Then $M_n \xrightarrow{d} M$ as $n \to \infty$, in $D[0,\infty)$, where M is a continuous q-dimensional Gaussian process with $\mathbb{E}[M(x)] = 0$ and covariances

$$\mathbb{E}[M(x)M'(y)] = \Sigma(x), 0 \le x \le y < \infty.$$

Chapter 3

Recursive trees and forests

3.1 Recursive trees

A *recursive tree* is a rooted tree whose vertices are labelled 1 to n, where n is the number of vertices, in such a way that the root is labelled 1, and every path beginning at the root has labels in increasing order.

If v is not the root, the *parent* of v is the vertex closest to v on the path from the root to v. If v is the parent of w, then w is a *child* of v. The *degree* of a vertex is, as usual, the number of its neighbours; the *out-degree* of a vertex is the number of children it has. For every vertex apart from the root the out-degree is one less than the degree. A *leaf* is a vertex of degree 1.

A recursive tree can be written as a list $(a_2, \ldots, a_n)$, where a_i is the label of the parent of the vertex labelled i, for $i = 2, \ldots, n$. By the definition of recursive trees we have $1 \leq a_i < i$. On the other hand, for every list $(a_2, \ldots, a_n)$ with the property that $1 \leq a_i < i$ for $i = 2, \ldots, n$, we can construct a recursive tree. We do this by starting with a single vertex labelled 1. Then, for $i = 2, \ldots, n$, we include a vertex labelled i and attach it to the vertex labelled a_i. Different lists yield different recursive trees, and vice versa, so there is a bijection between recursive trees on n vertices and lists $(a_2, \ldots, a_n)$ where $1 \leq a_i < i$ for $i = 2, \ldots, n$. It follows that there are $(n-1)!$ recursive trees on n vertices. The 3! recursive trees on four vertices are pictured in Figure 3.1.

We let a *random recursive tree* on n vertices be a recursive tree chosen uniformly at random among all the recursive trees on n vertices. Thus every recursive tree on n vertices is chosen with probability $\frac{1}{(n-1)!}$. There is a simple graph process which produces random recursive trees $T_1, T_2, \ldots$, where T_k is a recursive tree on k vertices for $k \geq 1$. Let T_1 be a tree consisting of one vertex, with the label 1. For $k \geq 2$, we construct T_k from T_{k-1} by adding a

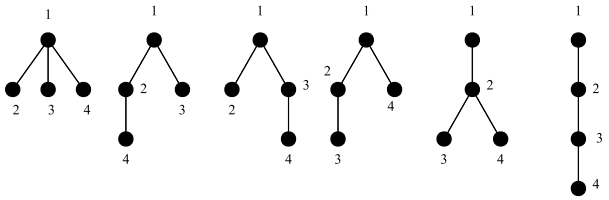

Figure 3.1: The six recursive trees of order 4

vertex labelled k, and attaching it by an edge to one of the vertices labelled 1 to $k-1$; this vertex is chosen uniformly at random, independently of the structure of T_{k-1}.

Equivalently, we may construct the random list $(a_2, \ldots, a_k)$, where a_i is chosen from the set $[i-1]$ uniformly at random, independently of the value of a_j for $j < i$. This procedure of choosing the list $(a_2, \ldots, a_k)$ ensures that the list is chosen uniformly at random from the set of lists $\{a_i\}_{i=2}^{k}$ with $1 \leq a_i < i$ for $i = 2, \ldots, k$. Because of the bijection between lists and recursive trees, it follows that the recursive trees produced by the graph process are also uniformly random. Many interesting results regarding various properties of random recursive trees, such as the height and independence number, appear in the survey [51] by Smythe and Mahmoud.

Recursive trees play an important role in the analysis of the minimum degree graph process, which we introduced in Section 1.6, and will examine closer in Chapter 4. To see why, let us consider the minimum degree graph process until the minimum degree becomes 1. Recall that $G_{\min}(n, 0)$ is the empty graph on n vertices, and that if $G_{\min}(n, m)$ contains isolated vertices, then $G_{\min}(n, m+1)$ is obtained from $G_{\min}(n, m)$ by choosing an isolated vertex v uniformly at random, and a vertex w uniformly at random from the set of vertices distinct from v, and adding an edge between v and w to the graph.

Let $H = H^{(n)}$ be the smallest value such that $G_{\min}(n, H)$ does not contain isolated vertices, and let $F^{(n)} = G_{\min}(n, H)$. Since $F^{(n)}$ is constructed by successively attaching an isolated vertex to some other vertex, no cycle can be formed, so $F^{(n)}$ must be a forest. Let C be any component in $F^{(n)}$, and suppose that C has k vertices. Since C is a tree, it contains $k-1$ edges, which we name e_1 to e_{k-1} in the order they were added in the graph process. Then e_1 connects two isolated vertices, creating an initial tree of order 2, while each of the edges $e_2, \ldots, e_{k-1}$ (provided that $k > 2$) causes the tree to grow by one vertex. Furthermore, whenever a vertex is added to the tree, the vertex to which the new vertex is connected is chosen uniformly at random.

Thus, C grows in the same way that we constructed random recursive trees.

If T is some unlabelled tree on k vertices, the probability that C is isomorphic to T is therefore equal to the probability that the random recursive tree T_k is isomorphic to T. Hence, the graph $F^{(n)}$ is a forest where every component has been constructed as a recursive tree. We therefore feel justified in calling $F^{(n)}$ a *recursive forest*. Note however, that the trees in $F^{(n)}$ are not recursive with regard to an initial labelling of the vertices, and our random recursive forests should not be confused with the perhaps more natural model introduced by Balińska, Quintas and Szymański [4], who define a recursive forest to be a forest on the vertex set $[n]$, where every tree is recursive with respect to the original labelling, and a random recursive forest is chosen uniformly at random from such forests.

In our discussion in Chapter 4 we will need to know the probability that an arbitrary component of order k in the random forest $F^{(n)}$ has exactly p leaves, or equivalently, the probability that the random recursive tree T_k has exactly p leaves. We will also need to know the number of components of order k in $F^{(n)}$, which we denote by $C_k^{(n)}$. We will solve these two problems in this chapter, but the main part of the chapter is devoted to showing that the random variables $C_1^{(n)}, C_2^{(n)}, \ldots$ tend to a Gaussian distribution as $n \to \infty$.

Let $e_{k,p}$ be the probability that T_k has precisely p leaves. If $k < 2$, we let $e_{k,p} = 0$ for all p. We define the bivariate generating function, related to $e_{k,p}$, as

$$E(u, v) = \sum_{k,p} e_{k,p} u^k v^p.$$

In Section 3.2 we will find the following exact expression for this generating function.

Theorem 3.1.

$$E(u, v) = u(1 - v)\left(\frac{1}{1 - ve^{u(1-v)}} + \log\left(1 - ve^{u(1-v)}\right) - \log(1 - v)\right) - 1. \tag{3.1}$$

The number of leaves, and more generally the degree distribution, of random recursive trees has also been considered by Mahmoud and Smythe [41] and Janson [25]. Unlike Theorem 3.1, which tells us the exact probabilities that a tree has a certain number of leaves, these authors consider the asymptotic properties of the degree sequence of T_k as $k \to \infty$, showing that the number of vertices of out-degree i for $i \geq 0$ have a multivariate normal distribution. The main theorem of [25] is given below.

Theorem 3.2 (Janson 2005)**.** *Let X_{ni} be the number of vertices of outdegree $i \geq 0$ in a random recursive tree with n vertices. As $n \to \infty$, $n^{-1}X_{ni} \to 2^{-i-1}$ a.s., and*

$$n^{-1/2}(X_{ni} - 2^{-i-1}n) \xrightarrow{d} V_i,$$

jointly for all $i \geq 0$, where the V_i are jointly Gaussian variables with means $\mathbb{E}V_i = 0$ and covariances $\sigma_{ij} = \mathrm{Cov}(V_i, V_j)$ given by the generating function

$$\sum_{i,j=0}^{\infty} \sigma_{ij} z^j w^i = \frac{2(1-z)^2(1-w)^2}{(2-z)(2-w)(2-zw)(3-z-w)}.$$

Theorem 3.2 was proved using the martingale central limit theorem of Section 2.4, Theorem 2.6. In this chapter, we will prove a similar theorem regarding the number of components of different orders in the random forest $F^{(n)}$. The approach we will use is similar to the one used in [25] to prove Theorem 3.2, but differs in some places, due to the nature of the processes.

Janson studied the random recursive tree process which goes on indefinitely. Proving an asymptotic result requires looking at the process as the number of steps tends to infinity. In the random forest, we start with n isolated vertices, and the process ends as soon as the isolated vertices have all been spent, which happens after a finite, random number of steps. Thus, proving an asymptotic result involves considering a sequence of processes $\left\{F^{(n)}\right\}_{n\geq 2}$ as $n \to \infty$. We will show that the approach used in [25] can be adapted to this situation, so that the following theorem can be proved.

Theorem 3.3. *Let C_{nk} be the number of components of order k in the random forest $F^{(n)}$. As $n \to \infty$, $n^{-1}C_{nk} \to \frac{k-1}{k}2^{-k}$ a.s., and $n^{-1/2}(C_{nk} - \frac{k-1}{k}2^{-k}n) \xrightarrow{d} W_k$, jointly for all $k \geq 1$, where the W_k are jointly Gaussian variables with means $\mathbb{E}W_k = 0$, and computable covariances $\mathrm{Cov}(W_i, W_j)$.*

By "computable", it is here meant that we can calculate $\mathrm{Cov}(W_i, W_j)$ explicitly for every choice of i and j. Moreover, we conjecture that

$$\mathrm{Cov}(W_i, W_j) = -\frac{ij(i-1)(j-1)}{2^{2+i+j}} + \frac{i(i-1)}{2^i}\delta_{ij}.$$

3.2 The number of leaves

In this section we will prove Theorem 3.1, namely that (3.1) holds. For $p \geq 2$, we define the monovariate generating function $E_p(z)$ by

$$E_p(z) = \sum_{k\geq p} e_{k,p} z^k.$$

We will first find an expression for $E_p(z)$. It will then be straightforward to sum over p and obtain (3.1).

Lemma 3.4. *For $p \geq 2$,*

$$E_p(z) = -\frac{z}{p(p-1)} + \sum_{i=2}^{p} \frac{(-iz)^{p-i}e^{iz}}{(p-i)!i^2}(i-1)(iz+p-i). \tag{3.2}$$

Proof. We say that a tree is a (k,p)-tree if it consists of k vertices, exactly p of which are leaves.

It is easy to see that $e_{k,1} = 0$ for all $k \geq 2$, $e_{k,k} = 0$ for all $k > 2$ and $e_{2,2} = 1$. Suppose that $k > 2$ and $2 \leq p < k$. A (k,p)-tree can either be constructed from a $(k-1,p)$-tree by attaching the next vertex to a leaf, or from a $(k-1,p-1)$-tree by attaching it to a non-leaf. Since no other way is possible, $e_{k,p}$ satisfies the recursion

$$e_{k,p} = \frac{k-p}{k-1}e_{k-1,p-1} + \frac{p}{k-1}e_{k-1,p}, \tag{3.3}$$

when $k > 2$.

We prove (3.2) by induction on p. Assume that $p > 2$. Multiplying (3.3) by $(k-1)z^k$ and summing over k, we get

$$\sum_k (k-1)e_{k,p}z^k = \sum_k ((k-p)e_{k-1,p-1} + pe_{k-1,p})z^k,$$

which can be written as

$$z^2\left(\frac{E_p(z)}{z}\right)' = z\left(zE_{p-1}(z)\right)' - pzE_{p-1}(z) + pzE_p(z).$$

This gives

$$zE_p'(z) - (1+pz)E_p(z) = z(1-p)E_{p-1}(z) + z^2E_{p-1}'(z),$$

which is a differential equation in $E_p(z)$. We solve it for $E_p(z)$ and obtain the recursive formula

$$E_p(z) = zE_{p-1}(z) + ze^{pz}\int E_{p-1}(z)e^{-pz}\left(p - \frac{p-1}{z}\right)dz, \tag{3.4}$$

for $p > 2$. We now need an expression for $E_2(z)$ as a basis for induction. A tree on two vertices must have two leaves, so $e_{2,2} = 1$. Since $e_{k,1} = 0$ for all $k \geq 2$, (3.3) implies that for $k > 2$,

$$e_{k,2} = \frac{2}{k-1}e_{k-1,2}.$$

It is then easy to derive that $e_{k,2} = \frac{2^{k-2}}{(k-1)!}$, and therefore that

$$E_2(z) = \frac{z}{2}\left(e^{2z} - 1\right).$$

Together with the fact that the coefficient of z in $E_p(z)$ is always zero, (3.2) now follows by induction, using (3.4). □

With this information we are able to calculate the generating function $E(u, v)$. By standard methods one gets that

$$\sum_{p\geq 2} \frac{v^p}{p(p-1)} = v + (1-v)\log(1-v).$$

Moreover,

$$\begin{aligned}
& \sum_p \sum_{i=2}^{p} \frac{(-iu)^{p-i}e^{iu}}{(p-i)!i^2}(i-1)(iu+p-i)v^p \\
= & \sum_{i\geq 2} \frac{(i-1)e^{iu}}{i^2} v^i \sum_{j\geq 0} \frac{(-iu)^j}{j!}(iu+j)v^j \\
= & \sum_{i\geq 2} \frac{(i-1)e^{iu}}{i^2} v^i iu(1-v)e^{-iuv} \\
= & u(1-v)\left(\frac{1}{1-ve^{u(1-v)}} + \log(1-ve^{u(1-v)}) - 1\right).
\end{aligned}$$

Putting this together,

$$\begin{aligned}
E(u,v) &= \sum_p E_p(u)v^p \\
&= \sum_p \left(-\frac{u}{p(p-1)} + \sum_{i=2}^{p} \frac{(-iu)^{p-i}e^{iu}}{(p-i)!i^2}(i-1)(iu+p-i)\right) v^p \\
&= -u(v + (1-v)\log(1-v)) \\
&\quad + u(1-v)\left(\frac{1}{1-ve^{u(1-v)}} + \log(1-ve^{u(1-v)}) - 1\right).
\end{aligned}$$

This completes the proof of Theorem 3.1. Table 3.1 contains the values of $e_{k,p}$ for $k \leq 9$. These can be obtained from (3.1), or they can simply be calculated using the recursive formula (3.3). Table 3.2 gives the number of recursive trees with a given number of leaves. Recently David Callan discovered the number triangle in Table 3.2 in a different context (Sequence A120434 in [50]). The entry in row k and column p is also the number of permutations of $[k-1]$ with exactly $p-2$ *big descents*, where a big descent in a permutation

$k\backslash p$	2	3	4	5	6	7	8
2	1						
3	1						
4	$\frac{2}{3}$	$\frac{1}{3}$					
5	$\frac{1}{3}$	$\frac{7}{12}$	$\frac{1}{12}$				
6	$\frac{2}{15}$	$\frac{11}{20}$	$\frac{3}{10}$	$\frac{1}{60}$			
7	$\frac{2}{45}$	$\frac{131}{360}$	$\frac{19}{40}$	$\frac{41}{360}$	$\frac{1}{360}$		
8	$\frac{4}{315}$	$\frac{473}{2520}$	$\frac{151}{315}$	$\frac{359}{1260}$	$\frac{11}{315}$	$\frac{1}{2520}$	
9	$\frac{1}{315}$	$\frac{179}{2240}$	$\frac{2399}{6720}$	$\frac{4211}{10080}$	$\frac{149}{1120}$	$\frac{61}{6720}$	$\frac{1}{20160}$

Table 3.1: The probability that a random recursive tree with k vertices has precisely p leaves

$k\backslash p$	2	3	4	5	6	7	8
2	1						
3	2						
4	4	2					
5	8	14	2				
6	16	66	36	2			
7	32	262	342	82	2		
8	64	946	2416	1436	176	2	
9	128	3222	14394	16844	5364	366	2

Table 3.2: The number of recursive trees with k vertices and p leaves

$(x_1, x_2, \ldots, x_n)$ is a position i such that $x_i - x_{i+1} \geq 2$. Interestingly, if a leaf is defined to be a vertex of out-degree 0, rather than a vertex of degree 1, then (see page 25 in Stanley [52]) the number of recursive trees on k vertices with p leaves equals the number of permutations of $[k-1]$ with p descents, with the last entry in the permutation always considered a descent.

We close this section by also mentioning the expectation and variance of the number L_k of leaves in T_k.

$$\begin{aligned} \mathbb{E}[L_k] &= \frac{k}{2} + \frac{1}{k-1}, \\ \mathrm{Var}[L_k] &= \frac{(k-3)(k^2+k+4)}{12(k-1)^2}. \end{aligned}$$

3.3 A random recursive forest

For the rest of the chapter we will consider the random forest $F^{(n)}$, which is obtained by running the minimum degree graph process $G_{\min}(n, m)$ until no isolated vertex is left. Recall that $H = H^{(n)}$ is the smallest value such that $G_{\min}(n, H)$ does not contain isolated vertices. If $0 \leq m \leq H$, we let $F_m^{(n)} = G_{\min}(n, m)$. Thus $F^{(n)} = F_H^{(n)}$. For $m > H$, we let $F_m^{(n)} = F_H^{(n)}$. It was proved in [30] that $H = n \log 2 + o(n)$ a.a.s. Later we will prove that H tends to a normal distribution with standard deviation $\Theta(\sqrt{n})$ when n tends to infinity.

From now on we will suppress the variable n, so we write $F = F^{(n)}$, $F_m = F_m^{(n)}$ and $H = H^{(n)}$. However, it should not be forgotten that these random variables, along with most random variables to be introduced later in this chapter, do indeed depend on the number of vertices, n.

Let $t = \frac{m}{n}$. We will imagine that t represents the time, so that we start with an empty graph at time $t = 0$, which gradually fills up as the time goes. As noted above, the last isolated vertex is likely to disappear when $t = \log 2$, at which point the process becomes stagnant.

For $m \geq 0$, $k \geq 1$, let C_{mk} be the random variable denoting the number of components of order k in F_m. Let C_k be the random variable denoting the number of components of order k in the random forest F. We will calculate the expected value of C_{mk}, and show that the random variables C_{mk} are concentrated around their expectations. For this we will use the differential equation method of Section 2.3. First we have to make sure that the largest component is not too large.

Lemma 3.5. *The largest tree in $F^{(n)}$ has a.a.s. $O(\log n)$ vertices.*

Proof. We have to prove that there is a positive constant c, such that there a.a.s. is no tree of order $c \log n$ or greater.

When an edge (v, w) is added to the graph, we can think of it the way that we first choose a vertex v of minimum degree, and then let v choose the vertex w randomly from the remaining vertices. Then v is the *choosing* vertex, while w is the *chosen* vertex. Set $k = \lceil c \log n \rceil - 1$, and suppose that there is a component of order at least $k + 1$. Let E be the set of edges in $F_m^{(n)}$. Then there is a set of edges $E' = \{e_1, \ldots, e_k\} \subset E$ with the following property:

For $i = 1, \ldots, k$, let $e_i = (v_i, w_i)$, where v_i is the choosing vertex of the edge. Then for every $i = 2, \ldots, k$, $w_i \in \{v_1, v_2, \ldots, v_{i-1}, w_1\}$.

Let $E'' \subset E$ be any subset of E with $|E''| = k$. The probability that E'' satisfies the above property is at most

$$\frac{2}{n-1}\frac{3}{n-1}\cdots\frac{k-1}{n-1} = \frac{(k-1)!}{(n-1)^{k-2}}.$$

Since $|E| = nh_1 + o(n)$ a.a.s., with $h_1 = \log 2 \approx 0.69$, there are about $\binom{nh_1}{k}$ ways to choose a set of k edges from E. Hence the probability that there is a set E' as described above, is bounded from above by

$$\binom{0.7n}{k}\frac{k!}{(n-1)^{k-2}} \le \frac{(0.7n)^k}{k!}\frac{k!}{(n-1)^{k-2}} \sim n^2 0.7^{k-2},$$

which tends to 0 for a sufficiently large c. □

Theorem 3.6. *Let $0 \le t < \log 2$. Then $C_{mk} = \beta_k(t)n + o(n)$ a.a.s., where*

$$\beta_k(t) = \frac{1}{k}(1 - e^{-t})^{k-1}((k+1)e^{-t} - 1). \tag{3.5}$$

Furthermore a.a.s.

$$C_k = \frac{k-1}{k2^k}n + o(n). \tag{3.6}$$

Proof. The proof relies on Theorem 2.5. We first find an expression for the expected amount of change from C_{mk} to $C_{m+1,k}$. When an edge is added to the graph, the first end of the edge is always an isolated vertex. We therefore always lose a component of order 1. If the other end is in a component of order k, we lose one component of order k, and if it is in a component of order $k-1$, we gain one component of order k. The probabilities of these two events are $\frac{kC_{mk}-\delta_{k1}}{n-1}$ and $\frac{(k-1)C_{m,k-1}-\delta_{k2}}{n-1}$, respectively. Hence, for $k \ge 1$,

$$\mathbb{E}\left[C_{m+1,k} - C_{m,k} | F_m\right] = -\delta_{k1} - \frac{kC_{mk}}{n} + \frac{(k-1)C_{m,k-1}}{n} + o(1). \tag{3.7}$$

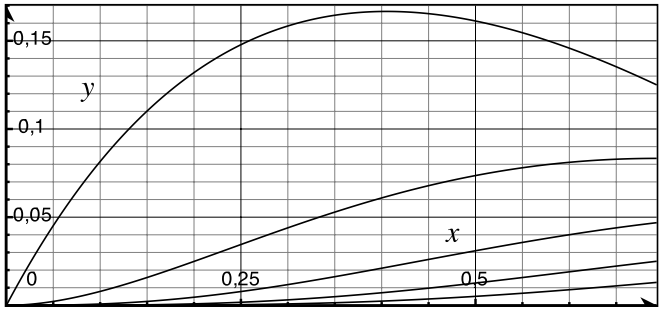

Figure 3.2: The number of trees of order 2 and greater in $F_{tn}^{(n)}$, divided by n

Furthermore it is clear that $C_{mk} \leq n$, and that $|C_{m+1,k} - C_{m,k}| \leq 2$, for all $k \geq 1$ and $m \geq 0$. Thus, by Theorem 2.5, there are functions $\beta_k(t)$ such that a.a.s. $C_{mk} = \beta_k(t)n + o(n)$, and such that the $\beta_k(t)$'s satisfy the differential equations

$$\frac{d}{dt}\beta_k(t) = -\delta_{k1} - k\beta_k(t) + (k-1)\beta_{k-1}(t), \tag{3.8}$$

which are obtained from (3.7). Furthermore, we have the boundary conditions

$$\beta_1(0) = 1, \tag{3.9}$$
$$\beta_k(0) = 0, \ k \geq 2. \tag{3.10}$$

Since (3.8) is linear, it satisfies a Lipschitz condition in a suitable domain D, so all the requirements of Theorem 2.5 are satisfied. Thus, we just have to find the functions $\beta_k(t)$. For $k = 1$, (3.8) becomes $\frac{d}{dt}\beta_1(t) = -1 - \beta_1(t)$, which, together with (3.9), gives us

$$\beta_1(t) = 2e^{-t} - 1. \tag{3.11}$$

We can now verify that (3.11) is the same as (3.5) with $k = 1$, and we proceed to prove by induction that (3.5) holds in the general case.

For $k \geq 2$, using standard techniques to solve the differential equation (3.8) with respect to $\beta_k(t)$, we obtain the recursive formula

$$\beta_k(t) = (k-1)e^{-kt}\int \beta_{k-1}(t)e^{kt}dt. \tag{3.12}$$

Using this recursive formula, and (3.10), it follows by induction that

$$\beta_k(t) = \frac{1}{k}\left(1 - e^{-t}\right)^{k-1}\left((k+1)\,e^{-t} - 1\right).$$

This implies the first part of the theorem, by Theorem 2.5.

To prove the second part of the theorem, we recall that the last isolated vertex a.a.s. disappears after $n \log 2 + o(n)$ steps, so we can therefore calculate the number of components of order k in G by setting $t = \log 2$ in (3.5), thereby obtaining (3.6). □

Theorem 3.6 gives us a fairly good estimate on how many trees there are of a given order at different stages of the random graph process. If we look closer, we see that the number of components of any given order is, in fact, normally distributed, and that if we consider the number of components of several different orders at the same time, they have a multivariate normal distribution. Theorem 3.3 states this in precise terms for the graph F, the random forest which is the final graph of the process. We will prove the following theorem, which holds for any value $0 \le t < \log 2$, and then derive Theorem 3.3 from it. For technical reasons we will consider V_{mk}, the number of vertices in components of order k in F_m, rather than C_{mk}. For all $k \ge 1$, we clearly have $V_{mk} = kC_{mk}$.

Theorem 3.7. *Let $0 \le t < \log 2$, and $m = tn$. As $n \to \infty$, $n^{-1}V_{mk} \to k\beta_k(t)$, and*

$$n^{-1/2}(V_{mk} - k\beta_k(t)n) \xrightarrow{d} W_{tk},$$

jointly for all $k \ge 1$, where the W_{tk} are jointly Gaussian variables with means $\mathbb{E}[W_{tk}] = 0$ and computable covariances $c_{ij}(t) = \mathrm{Cov}(W_{ti}, W_{tj})$, where $c_{ij}(t)$ are continuous functions of t for all $i, j \ge 1$.

Theorem 3.7 complements Theorem 3.3, in that it tells how F_{tn} looks strictly before the end of the process. Note that, although we obtained the expected number of components of order k in Theorem 3.3 by simply letting $t \to \log 2$ in $\beta_k(t)$, this does not work for the covariances $c_{ij}(t)$. Indeed, we will observe later in this chapter that $\lim_{t\to\log 2} c_{11}(t) \neq 0$, whereas the covariance $\mathrm{Cov}(V_1, V_1) = \mathrm{Var}(V_1)$ must necessarily be 0, since there are no components of order 1 in the final graph. This is explained by the fact that the stopping time H is a random variable itself, whose variance must be taken into account.

3.4 Martingale

We will prove Theorem 3.7 using Theorem 2.6. In order to use this theorem, we have to express our random variables as martingales. The first problem we encounter is that for every t, there is an infinite number of random variables $V_{tn,1}, V_{tn,2}, V_{tn,3}, \ldots$. In order to use the martingale central limit theorem, we must reduce this to a finite number of random variables, and we do this by capping the largest value at some point.

Let $q > 1$ be a fixed integer. Let $Z_{mk} = V_{mk}$ for $k = 1, \ldots, q-1$, and $Z_{mq} = \sum_{k \geq q} V_{mk}$. Thus, if $k < q$, then Z_{mk} is the number of vertices in components of order k, whereas Z_{mq} is the number of vertices in components of order at least q. We always have $\sum_{k=1}^{q} Z_{m,k} = n$. Let $Z_m = [Z_{m1}, \ldots, Z_{mq}]'$ be the corresponding stochastic q-dimensional column vector. The process $\{Z_m\}_m$ is a Markov process, since the distribution of Z_m only depends on Z_{m-1}. The process is not a martingale, but we will define a transformation of the vector Z_m, such that the resulting process is a martingale.

Let $T = \{t_{ij}\}_{i,j=0}^{q}$ be the matrix defined by

$$t_{ij} = \begin{cases} 1 - \frac{i}{n-1} & \text{if } i = j < q, \\ 1 + \frac{1}{n-1} & \text{if } i = j = q, \\ \frac{i}{n-1} & \text{if } i = j+1, \\ 0 & \text{otherwise.} \end{cases}$$

We see that

$$\mathbb{E}\left[Z_m | F_{m-1}\right] = T Z_{m-1} - \mathbf{e}_1. \tag{3.13}$$

If we let

$$A = (n-1)(T - I), \tag{3.14}$$

then

$$A = \begin{bmatrix} -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 2 & -2 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 3 & -3 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2-q & 0 & 0 \\ 0 & 0 & 0 & \cdots & q-1 & 1-q & 0 \\ 0 & 0 & 0 & \cdots & 0 & q & 1 \end{bmatrix}.$$

The exponential function of a matrix is defined by a power series: $e^{tA} = \sum_{i=0}^{\infty} \frac{(tA)^j}{j!}$.

Lemma 3.8. *If $m = tn$, then $T^m \sim e^{tA}$.*

Proof. Both A and T are lower triangular matrices with q distinct eigenvalues, and λ is an eigenvalue of A if and only if $1 + \frac{\lambda}{n}$ is an eigenvalue of T. Let $\lambda_1, \ldots, \lambda_q$ be the eigenvalues of A, and for $i = 1, \ldots, q$, let $\lambda_i' = 1 + \frac{\lambda_i}{n}$. Let D be the diagonal matrix with entries $\lambda_1, \ldots, \lambda_q$, and let D' be the diagonal matrix with entries $\lambda_1', \ldots, \lambda_q'$. Let $\mathbf{v}_i$ be a vector such that $A\mathbf{v}_i = \lambda_i \mathbf{v}_i$. Then $T\mathbf{v}_i = \left(I + \frac{1}{n}A\right)\mathbf{v}_i = \left(1 + \frac{\lambda_i}{n}\right)\mathbf{v}_i = \lambda_i'\mathbf{v}_i$. Hence A and T have the same eigenvectors. It follows that there is a matrix Q such that $A = QDQ^{-1}$ and $T = QD'Q^{-1}$. Since $\lambda_i'^{tn} \sim e^{t\lambda_i}$, we have $D'^{tn} \sim e^{tD}$, and so $T^m = T^{tn} = QD'^{tn}Q^{-1} \sim Qe^{tD}Q^{-1} = e^{tA}$. □

In order to find a martingale, we normalise the random vector Z_m. The first step is to find the expectation of Z_m. This is

$$E_m := n(T^m(I - A^{-1}) + A^{-1})\mathbf{e}_1.$$

By Lemma 3.8, $E_m \sim n(e^{tA}(I - A^{-1}) + A^{-1})\mathbf{e}_1$. To see that $\mathbb{E}[Z_m] = E_m$, note that $E_0 = n\mathbf{e}_1 = Z_0$, and that the two equations

$$\mathbb{E}[Z_m] = T\mathbb{E}[Z_{m-1}] - \mathbf{e}_1$$

and

$$E_m = TE_{m-1} - \mathbf{e}_1 \tag{3.15}$$

hold for $m \geq 1$. By induction it follows that $\mathbb{E}[Z_m] = E_m$ for all $m \geq 0$. We get a proper normalisation of Z_m to a martingale by defining

$$S_m = T^{-m}(Z_m - E_m).$$

Equations (3.13) and (3.15) imply that

$$\begin{aligned} \mathbb{E}[S_m - S_{m-1}|S_{m-1}] &= T^{-m}((\mathbb{E}[Z_m|Z_{m-1}] - TZ_{m-1}) - (E_m - TE_{m-1})) \\ &= T^{-m}(-\mathbf{e}_1 - (-\mathbf{e}_1)) = 0, \end{aligned}$$

so S_m is indeed a martingale.

3.5 Quadratic variation

We now calculate the quadratic variation of the q-dimensional martingale $\{S_m\}_m$, as defined in Section 2.4. If $\mathbf{v}$ is a q-dimensional vector, we let the square of the vector be the $q \times q$ matrix $\mathbf{v}^2 = \mathbf{v}\mathbf{v}'$, where $\mathbf{v}'$ is the transpose of $\mathbf{v}$.

For $m \geq 1$, we let $X_m = S_m - S_{m-1}$ be the martingale differences. The quadratic variation of S_m is $U_m^2 = \sum_{i=1}^m X_m^2$. Thus

$$\mathbb{E}\left[U_m^2 - U_{m-1}^2 | F_{m-1}\right] = \mathbb{E}\left[X_m^2 | F_{m-1}\right]. \tag{3.16}$$

We will now find an expression for the right hand side of (3.16), and then use Theorem 2.5 to prove that there is a matrix $\Sigma(t)$ such that a.a.s. $U_m^2 = \Sigma(t)n + o(n)$, where the $o(n)$ is a matrix, all of whose entries are $o(n)$.

Let $\Delta_m = Z_m - Z_{m-1}$. Then

$$Z_m - TZ_{m-1} = (I - T)Z_{m-1} + \Delta_m,$$

and

$$S_m - S_{m-1} = T^{-m}((I - T)Z_{m-1} + \Delta_m + \mathbf{e}_1)T'^{-m}.$$

We let $P_m = \frac{1}{n}Z_m$. Let $P_{m,i}$ be the ith entry of P_m. If $1 \leq i < q$, then $P_{m,i}$ is the probability that a vertex in a component of order i is chosen as the second vertex in step m; $P_{m,q}$ is the probability that a vertex in a component of order at least q is chosen. Using (3.14), the above equation can be written as

$$S_m - S_{m-1} = T^{-m}(-AP_{m-1} + \Delta_m + \mathbf{e}_1).$$

If Y is any random variable, we define, for short, $\mathbb{E}^*[Y] = \mathbb{E}[Y|F_{m-1}]$. Then

$$\mathbb{E}^*[\Delta_m] = \mathbb{E}^*[Z_m - Z_{m-1}] = TZ_{m-1} - \mathbf{e}_1 - Z_{m-1} = AP_{m-1} - \mathbf{e}_1.$$

Thus $\mathbb{E}^*[X_m^2]$ is given by

$$\begin{aligned}
& \mathbb{E}^*[(S_m - S_{m-1})(S_m - S_{m-1})'] \\
= \; & T^{-m}\mathbb{E}^*\left[(-AP_{m-1} + \Delta_m + \mathbf{e}_1)(-AP_{m-1} + \Delta_m + \mathbf{e}_1)'\right] T'^{-m} \\
= \; & T^{-m}\Big(\mathbb{E}^*[\Delta_m\Delta_m'] + \mathbb{E}^*[\Delta_m](-P_{m-1}'A' + \mathbf{e}_1) \\
& +(-AP_{m-1} + \mathbf{e}_1)\mathbb{E}^*[\Delta_m'] + (-AP_{m-1} + \mathbf{e}_1)(-P_{m-1}'A + \mathbf{e}_1)\Big)T'^{-m} \\
= \; & T^{-m}\Big(\mathbb{E}^*[\Delta_m\Delta_m'] - (AP_{m-1} - \mathbf{e}_1)(P_{m-1}'A' - \mathbf{e}_1) \\
& -(AP_{m-1} - \mathbf{e}_1)(P_{m-1}'A' - \mathbf{e}_1) + (AP_{m-1} - \mathbf{e}_1)(P_{m-1}'A - \mathbf{e}_1)\Big)T'^{-m} \\
= \; & T^{-m}\left(\mathbb{E}^*[\Delta_m\Delta_m'] - (AP_{m-1} - \mathbf{e}_1)\left(P_{m-1}'A' - \mathbf{e}_1'\right)\right)T'^{-m}.
\end{aligned}$$

The entry in the ith row and jth column of $\mathbb{E}^*[\Delta_m \Delta_m']$ is $\mathbb{E}[(Z_{m,i} - Z_{m-1,i})(Z_{m,j} - Z_{m-1,j})]$. Suppose that a vertex in a component of order k is chosen in the mth step. If $k = i$, we lose a component of order i, and if $k = i - 1$, we gain a component of order i. Since we also lose a component of order 1 in every step, we get $Z_{m,i} - Z_{m-1,i} = -\delta_{i1} - i\delta_{ki} + i\delta_{k,i-1}$ if $i < q$. If $i = q$, then we gain a component of order q if $k = q - 1$. If $k \geq q$, the number of vertices in components of order at least q increases by 1. Thus, $Z_{m,q} - Z_{m-1,q} = q\delta_{k,q-1} + I_{k \geq q}$. This can be expressed as $Z_{m,i} - Z_{m-1,i} = a_{ik} - \delta_{i1}$ for $1 \leq i \leq q$, where a_{ik} is the entry in the ith row and the kth column in A. Thus

$$(Z_{m,i} - Z_{m-1,i})(Z_{m,j} - Z_{m-1,j}) = \sum_{k=1}^{q} P_{m-1,k} a_{ik} a_{jk}.$$

This implies that

$$\mathbb{E}^*[\Delta_m \Delta_m'] = (A - J) D_{P_{m-1}} (A' - J'),$$

where J is the $q \times q$ matrix $\{\delta_{i1}\}_{i,j=0}^{q}$, and $D_{P_{m-1}}$ is the diagonal matrix whose ith entry is $P_{m-1,i}$. We note that $D_{P_{m-1}} J' = P_{m-1} \mathbf{e}_1'$, $J D_{P_{m-1}} = \mathbf{e}_1 P_{m-1}'$ and $J D_{P_{m-1}} J' = \mathbf{e}_1 \mathbf{e}_1'$. Hence

$$\mathbb{E}\left[U_m^2 - U_{m-1}^2 | F_{m-1}\right] = \mathbb{E}\left[X_m^2 | F_{m-1}\right] = T^{-m} A (D_{P_{m-1}} - P_{m-1} P_{m-1}') A' T'^{-m}.$$

By Lemma 3.8, this can be written as

$$\mathbb{E}\left[U_m^2 - U_{m-1}^2 | F_{m-1}\right] = e^{-tA} A (D_{P_{m-1}} - P_{m-1} P_{m-1}') A' e^{-tA'} + o(1). \quad (3.17)$$

Let

$$\gamma_k(t) = \begin{cases} \beta_k(t) & \text{if } 1 \leq k < q, \\ \sum_{i \geq q} \beta_i(t) & \text{if } k = q. \end{cases}$$

Theorem 3.6 implies that the kth entry in P_{m-1} is a.a.s. $\gamma_k(t) + o(1)$ for $k = 1, \ldots, q$. By (3.17)

$$\mathbb{E}\left[U_m^2 - U_{m-1}^2 | F_{m-1}\right] = e^{-tA} A (\{\delta_{ij}\gamma_i(t)\}_{i,j=1}^{q} - \{\gamma_i(t)\gamma_j(t)\}_{i,j=1}^{q}) A' e^{-tA'} + o(1). \quad (3.18)$$

Let $\Xi(t)$ be the matrix on the right hand side of (3.18) without the $o(1)$ term, and let $\xi_{ij}(t)$ be the entry in the ith row and jth column of $\Xi(t)$. It follows then from Theorem 2.5 that there are functions $\sigma_{ij}(t)$ such that a.a.s. $U_m^2 = \{\sigma_{ij}(t)n + o(n)\}_{i,j=0}^{q}$, which satisfy

$$\frac{d}{dt}\sigma_{ij}(t) = \xi_{ij}(t) \quad (3.19)$$

with the boundary conditions $\sigma_{ij}(0) = 0$. Let $\Sigma(t) = \{\sigma_{ij}(t)\}_{i,j=0}^{q}$. Then the above statement can be formulated as

$$n^{-1}U_m^2 \xrightarrow{p} \Sigma(t).$$

Let us now define $M_n(t) = \frac{1}{\sqrt{n}}S_{\lfloor tn \rfloor}$. Since $\{S_m\}_{m\geq 0}$ is a martingale, $\{M_n(t)\}_{t\geq 0}$ is also a martingale. Then the quadratic variation of $M_n(t)$ is

$$\begin{aligned}
[M_n, M_n]_t &= \sum_{0<s\leq t} \Delta M_n(s)\Delta M_n(s)' \\
&= \frac{1}{n}\sum_{m=0}^{\lfloor tn \rfloor}(S_m - S_{m-1})(S_m - S_{m-1})' \\
&= \frac{1}{n}\sum_{m=0}^{\lfloor tn \rfloor} X_m^2 = \frac{1}{n}U_{\lfloor tn \rfloor}^2 \xrightarrow{p} \Sigma(t).
\end{aligned}$$

Theorem 2.6 then implies that $M_n \xrightarrow{d} M$, where M is a Gaussian process with $\mathbb{E}M(t) = 0$ and with covariance matrix $\mathbb{E}M(t)M'(t') = \Sigma(t)$ for $0 \leq t \leq t' < \log 2$.

Since

$$n^{-1/2}(Z_m - E_m) = n^{-1/2}T^m S_m \sim e^{tA}M(t),$$

we conclude that Z_m tends to a Gaussian process with mean $(e^{tA}(I - A^{-1}) + A^{-1})\mathbf{e}_1$ and covariance matrix

$$\{c'_{ij}(t)\} = e^{tA}\Sigma(t)e^{tA'} \tag{3.20}$$

when $n \to \infty$. Since $V_{mk} = Z_{mk}$ for $1 \leq k < q$, this implies that the variables $V_{m1}, \ldots, V_{m,q-1}$ tend to a jointly normal distribution. Since q is arbitrary, we conclude that this holds for all the variables $V_{m1}, V_{m2}, \ldots$.

Moreover, since $\xi_{ij}(t)$ is given for all i and j by (3.18), the covariances $c'_{ij}(t)$ can be explicitly calculated by (3.19) and (3.20). Since e^{tA} is a lower triangular matrix, $c'_{ij}(t)$ does not depend on any entries in $\Sigma(t)$ which are below the ith row or to the right of the jth column. Similarly one can see from (3.18) that the expression for $\xi_{ij}(t)$ does not contain $\gamma_k(t)$ for any $k > \max(i, j)$. Thus, $c'_{ij}(t)$ does not depend on the value of q as long as $q > \max(i, j)$, and the covariance $c_{ij}(t)$ in Theorem 3.7 equals $c'_{ij}(t)$, provided that q is chosen large enough. This constitutes a proof of Theorem 3.7.

3.6 The final stage

We will now deduce Theorem 3.3 from Theorem 3.7. The latter theorem states that the random variables $V_{tn,1}, V_{tn,2}, \ldots$ are jointly normal whenever $t < \log 2$. We want to show that this holds also at the end of the

process, which happens a.a.s. when $t = \log 2$. Recall that H was defined in Section 3.3 as the stopping time of the process, that is the number of edges added when the last isolated vertex disappears.

In Theorem 3.7 W_{tk} is defined for $k \geq 0$ and $0 \leq t < \log 2$. We extend the definition to $t = \log 2$, by letting $W_{\log 2,k}$ be jointly normal variables for $k \geq 1$, such that $\mathrm{Cov}(W_{\log 2,i}, W_{\log 2,j}) = c_{ij}(\log 2)$ and $c_{ij}(t)$ is as in Theorem 3.7. If the stopping time H were not random, but deterministically equal to $\lfloor n \log 2 \rfloor$, we could simply say that $W_k = W_{\log 2,k}$ with $\mathrm{Cov}(W_i, W_j) = c_{ij}(\log 2)$. However, since H is a random variable, this affects the variables W_k.

Let us consider the case $k = 1$, the isolated vertices. For $t < \log 2$, we have $n^{-1/2}(V_{m1} - (2e^{-t} - 1)n) \xrightarrow{d} W_{t1}$. Thus, for every x,

$$\lim_{n\to\infty} \mathbb{P}\left[n^{-1/2}(V_{m1} - (2e^{-t} - 1)n) \geq x\right] = \mathbb{P}\left[W_{t1} \geq x\right] \tag{3.21}$$

When the process goes on, V_{m1} decreases, but it will stop decreasing as soon as it becomes 0. However, as long as $V_{m1} > 0$, (3.21) still holds. We can use this to show that the stopping time H is also normally distributed.

For all m,

$$H > m \Leftrightarrow V_{m,1} > 0.$$

Using the results of the previous section, we can calculate that if $t < \log 2$, then $\mathrm{Var}(W_{t,1}) = -1 + 6e^{-t} - (5 + 4t)e^{-2t}$, so $\mathrm{Var}(W_{\log 2,1}) = \frac{3}{4} - \log 2$. Thus $W_{\log 2,1} \sim \mathcal{N}(0, \frac{3}{4} - \log 2)$.

Let $\eta = \frac{H - n\log 2}{\sqrt{n}}$ and $m' = m - n\log 2$. Then

$$\begin{aligned}
& \mathbb{P}\left[\eta > \frac{m - n\log 2}{\sqrt{n}}\right] = \mathbb{P}[H > m] = \mathbb{P}[V_{m,1} > 0] \\
= \;& \mathbb{P}\left[W_{\log 2,1} > -\sqrt{n}(2e^{-m/n} - 1)\right] = \mathbb{P}\left[W_{\log 2,1} > -\sqrt{n}(e^{-m'/n} - 1)\right] \\
= \;& \mathbb{P}\left[W_{\log 2,1} > -\sqrt{n}(1 - \frac{m'}{n} + \left(\frac{m'}{n}\right)^2 + \ldots - 1)\right] \\
\sim \;& \mathbb{P}\left[W_{\log 2,1} > \frac{m - n\log 2}{\sqrt{n}}\right].
\end{aligned}$$

Thus η tends to a normal variable with mean 0 and variance $\frac{3}{4} - \log 2$ as $n \to \infty$. Moreover, $\mathrm{Cov}(\eta, W_{\log 2,1}) = 1$.

We now define a new sequence of random variables $V'_{m,k}$, for $m = 0, 1, \ldots$ and $k = 1, 2, \ldots$, in the following way. As long as $m = 0$ or F_{m-1} contains at least one isolated vertex, $V'_{m,k} = V_{m,k}$. Suppose that F_m does not contain any isolated vertices. Let $n'_m = \sum_{k\geq 2} V_{m,k}$. Then we let d_m be a random

variable such that $d_m = k$ with probability $\frac{V_{m,k}}{n'_m}$ for $k = 2, 3, \ldots$, and we let $V'_{m+1,k} = -\delta_{k,1} - k\delta_{d_m,k} + k\delta_{d_m,k-1}$. Essentially $V'_{m,k}$ continues to evolve in the same way it did earlier, as if there still were isolated vertices available. As long as $m = n \log 2 + o(n)$, we have $n'_m = (1 + o(1))n$, so

$$\mathbb{E}\left[V'_{m+1,k} - V'_{m,k} | V'_m\right] = -\delta_{k1} - \frac{kV'_{m,k}}{n} + \frac{kV'_{m,k-1}}{n} + o(1). \tag{3.22}$$

Using the arguments in Section 3.5 we then find that if $m = n \log 2 + o(n)$, then

$$n^{-1/2}\left(V'_{m,k} - k\beta_k\left(\frac{m}{n}\right)n\right) \xrightarrow{d} W_{\log 2,k}.$$

Since $V'_{m,k} = \frac{k-1}{2^k}n + o(n)$ a.a.s., (3.22) implies that

$$\mathbb{E}\left[V'_{m+1,k} - V'_{m,k} | V'_m\right] = -\delta_{k1} - \frac{k(k-1)}{2^k} + \frac{k(k-2)}{2^{k-1}} + o(1) = \frac{k(k-3)}{2^k} + o(1),$$

for $k = 1, 2, \ldots$. We are interested in the values $V'_{H,k} = V_{H,k}$, so we have to determine how much $V'_{m,k}$ changes when m changes from $n \log 2$ to H. We have $H - \log 2 = \eta\sqrt{n}$, and so if we let $v_k = \frac{k(k-3)}{2^k}$, then

$$\mathbb{E}\left[V_{H,k} - V_{n\log 2,k} | F^{(\min(H, n\log 2))}\right] = v_k \eta \sqrt{n}.$$

It follows that

$$n^{-1/2}\left(V_{H,k} - \frac{k-1}{2^k}n\right) \xrightarrow{d} W_{\log 2,k} + v_k W_{\log 2,1},$$

so $W_k = W_{\log 2,k} + v_k W_{\log 2,1}$.

Thus, every W_k is a linear combination of $\{W_{\log 2,i}\}_i$, so we can conclude that the W_k's are jointly normal. As for the covariances, they can be calculated by

$$\begin{aligned} \operatorname{Cov}(W_i, W_j) &= \operatorname{Cov}(W_{\log 2,i} + v_i W_{\log 2,1}, W_{\log 2,j} + v_j W_{\log 2,1}) \\ &= \operatorname{Cov}(W_{\log 2,i}, W_{\log 2,j}) + v_j \operatorname{Cov}(W_{\log 2,i}, W_{\log 2,1}) \\ &\quad + v_i \operatorname{Cov}(W_{\log 2,1}, W_{\log 2,j}) + v_i v_j \operatorname{Var}(W_{\log 2,1}). \end{aligned} \tag{3.23}$$

The arguments in the previous section tell us how the covariances $\operatorname{Cov}(W_{\log 2,i}, W_{\log 2,j})$ can be calculated. Thus, by (3.23), we can explicitly calculate $\operatorname{Cov}(W_i, W_j)$ for any i and j. For small values of i and j the covariance satisfies the formula

$$\operatorname{Cov}(W_i, W_j) = -\frac{ij(i-1)(j-1)}{2^{2+i+j}} + \frac{i(i-1)}{2^i}\delta_{ij},$$

but in general this formula seems difficult to prove.

Chapter 4

The minimum degree graph process

In this chapter we will consider the minimum degree graph process more closely. The main goal of the chapter is to determine the point of the phase transition and to show that the process undergoes a double jump at this point.

For the definition of the minimum degree graph process we refer to Section 1.6. Recall that $G_{\min}(n,m)$ is the graph after m edges have been added. Let $t = m/n$. For $i \geq 1$, let H_i be the smallest number such that the minimum degree of $G_{\min}(n, H_i)$ is at least i. Kang et al. [30] proved that there are constants h_1, h_2 and h_3 such that a.a.s. $H_i = h_i n + o(n)$ for each $i = 1, 2, 3$. These constants are

$$\begin{aligned} h_1 &= \log 2 \approx 0.6931, \\ h_2 &= \log 2 + \log(1 + \log 2) \approx 1.2197, \\ h_3 &= \log(\log^2 2 + 2(1 + \log 2)(1 + \log(1 + \log 2))) \approx 1.7316. \end{aligned} \tag{4.1}$$

Recall that in the standard random graph process the graph becomes a.a.s. connected when the last isolated vertex disappears, which occurs a.a.s. when the number of edges is $n \log n/2 + O(n)$. In contrast, Kang et al. proved that the minimum degree graph process becomes a.a.s. connected already after a linear number, $h_3 n$, of edges have been added; that is roughly when the minimum degree reaches 3. Furthermore, when $h_2 < t < h_3$, the probability that $G_{\min}(n, tn)$ is connected is bounded away from 0 and 1; if the graph is disconnected in this period, it consists a.a.s. of a giant component containing $n - o(n)$ vertices and a number of isolated cycles. In Section 4.1 we will derive the exact limit probability for connectivity when $h_2 < t < h_3$.

The main part of the chapter is dedicated to the phase transition of the minimum degree graph process. We will show that $G_{\min}(n,m)$ undergoes

a phase transition similar to $G(n, m)$, albeit somewhat later, when roughly $0.86n$ edges have been added.

4.1 Connectivity

Kang et al. [30] found the connectivity threshold for the minimum degree graph process.

Theorem 4.1 (Kang, Koh, Ree, Łuczak 2006)**.** *Let the constants h_2 and h_3 be defined as in (4.1), and let $\rho_n(t)$ denote the probability that $G_{\min}(n, tn)$ is connected. Then for every constant $t \neq h_2$ the limit*

$$\rho(t) = \lim_{n\to\infty} \rho_n(t)$$

exists, and $\rho(t) = 0$ for $t < h_2$, while $\rho(t) = 1$ for $t \geq h_3$. If $t \in (h_2, h_3)$, then $0 < \rho(t) < 1$ such that

$$\rho^+ = \lim_{t\to h_2+} \rho(t) > 0$$

and

$$\lim_{t\to h_3-} \rho(t) = 1.$$

The function $\rho(t)$ is continuous everywhere except at $t = h_2$. Kang et al. moreover proved that when $h_2 < t < h_3$, the graph $G_{\min}(n, tn)$ a.a.s. consists of one giant component with $n - o(n)$ vertices, and possibly one or more isolated cycles. They found that the number of isolated cycles has asymptotically Poisson distribution with mean $\lambda(t)$, where

$$\lambda(t) = \sum_{k=2}^{\infty} \frac{(k-1)^k}{k!} \left(\frac{e^{h_3-t} - 1}{e^{h_3} - e^{h_2}}\right)^k \mathbb{P}[G_{\min}(k, k) \text{ is a cycle}].$$

Thus, the probability that $G_{\min}(n, tn)$ is connected equals asymptotically the probability that the graph contains no isolated cycles, which tends to $e^{-\lambda(t)}$ as $n \to \infty$. Therefore,

$$\rho(t) = e^{-\lambda(t)}$$

and

$$\rho^+ = e^{-\lambda(h_2)}.$$

Kang et al. did not calculate the exact values of $\rho(t)$ and ρ^+, so we will begin our discussion of the minimum degree graph process by calculating these.

Let C be a fixed Hamiltonian cycle in the complete graph K_k. Since there are $\frac{(k-1)!}{2}$ Hamiltonian cycles in K_k, we have

$$\mathbb{P}\left[G_{\min}(k, k) \text{ is a cycle}\right] = \frac{(k-1)!}{2} \mathbb{P}\left[G_{\min}(k, k) = C\right]. \tag{4.2}$$

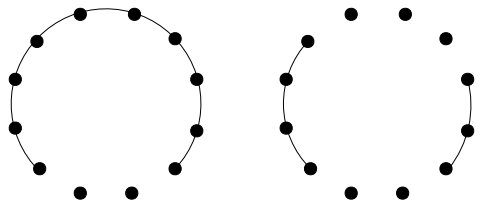

Figure 4.1: Cycles with gaps

Hence, if we let

$$h(t) = \frac{e^{h_3 - t} - 1}{e^{h_3} - e^{h_2}},$$

we can write

$$\lambda(t) = \sum_{k=2}^{\infty} \frac{(k-1)^k}{2k} h(t)^k \mathbb{P}\left[G_{\min}(k,k) = C\right]. \tag{4.3}$$

For a graph G we let $\iota(G)$ be the number of isolated vertices. Let $0 \leq i \leq k-2$ and $G_i = G_{\min}(k, k-i-1)$, and suppose that $G_i \subseteq C$ and $\iota(G_i) = i$. Then G_i consists of a path P of length $k-i-1$, which is a part of the Hamiltonian cycle C, and i isolated vertices. Let $f_{k,i}$ be the probability that $G_{\min}(k,k) = C$ in this case; that is

$$f_{k,i} = \mathbb{P}\left[G_{\min}(k,k) = C | G_i \subseteq C \wedge \iota(G_i) = i\right].$$

The graphs in Figure 4.1 are examples of the graph process for $k = 12$. The vertices are aligned according to the cycle C, which is the cycle we aim for. In the graph to the left in Figure 4.1, there are two isolated vertices, so the probability that this graph ends up as the cycle C is $f_{k,2}$.

Let us now consider the graph to the right in Figure 4.1. Here there are two gaps: one with two isolated vertices, and another with three. To determine the probability that this graph ends up as C, we must realise that we can handle each of the gaps essentially independently of the other.

Suppose more generally that $G = G_{\min}(k,t) \subseteq C$ has two gaps, as the graph to the right in Figure 4.1, such that upper gap U has i isolated vertices, and the lower gap L has j isolated vertices. Let G^U be the graph obtained by filling the lower gap in G with edges in such a way that $G^U \subseteq C$ and G^U only contains one gap. We now proceed with the minimum degree graph process on G^U, until either the remaining gap has been filled or until an edge which does not belong to C is added. Suppose u is the number of steps until the process stops, and let $(v_1, v_1'), \ldots, (v_u, v_u')$ be the edges added, where the v_m's are the vertices chosen first, and the v_m''s are the vertices chosen second

at every step. The probability that this process ends up with the upper gap being filled is, by definition, $f_{k,i}$.

Similarly, we let G^L be the graph obtained by filling the upper gap in G, and we proceed with the minimum degree graph process on G^L. We let the process continue until either the gap has been filled, or an edge which is not a part of the cycle C is added. We let l be the number of steps until this process stops, and we let $(w_1, w_1'), \dots, (w_l, w_l')$ be the edges added, where the w_m's are the first, and the w_m''s the second vertices chosen at every step. The probability that this process ends with the lower gap being filled is $f_{k,j}$.

We now go back to the original graph G, which contains the two gaps U and L. By considering the gaps separately above, we have got two randomly chosen sequences of edges, $V = ((v_1, v_1'), \dots, (v_u, v_u'))$ and $W = ((w_1, w_1'), \dots, (w_l, w_l'))$. We now proceed as follows, starting from the graph G. We choose the upper or the lower gap at random with probabilities proportional to the number of isolated vertices in each gap. If the upper gap is chosen, we add the first unused edge in V to G, and if the lower gap is chosen, we add the first unused edge in W to G. We continue like this until there are no more isolated vertices in the graph, or until an edge which does not form a part of C is added. After that we choose gaps with probabilities proportional to the number of vertices of degree 1. We then continue the process until either an edge which is not contained in C is added or until both gaps are filled. The event that both gaps are filled can only happen if both the upper process and the lower process were successful, so the probability is $f_{k,i}f_{k,j}$. Although we have chosen the edges with a different procedure than in the ordinary minimum degree graph process, the probability of a particular edge being chosen at any one step is unchanged. Hence we can conclude that the probability that both gaps in G are filled is $f_{k,i}f_{k,j}$.

Using this observation we can obtain a recursive formula for $f_{k,i}$. We again consider G_i, which consists of a path P which forms a part of C, and i isolated vertices which are consecutive vertices in C. There are then $i+1$ edges which are contained in C but not in P. We distinguish between the *internal* edges of the gap, both of whose endpoints are isolated in G_i, and *external* edges, which are incident to the path P. There are two external edges and $i-1$ internal edges.

Let us first consider the probability that one of the external edges is added in the next step. Let $e = vw$ be an external edge, where v is an isolated vertex in G_i and w is an endpoint of the path P. For e to be added, the first vertex chosen in the next step must be v, and the second must be w. The probability of the former event is $\frac{1}{i}$, and the probability of the latter is $\frac{1}{k-1}$. Since there are two external edges, the probability that one of them

is added at the next step is $\frac{2}{i(k-1)}$.

Let now $e = vw$ be an internal edge, where both v and w are isolated vertices. The probability that v is chosen as the first vertex and w as the second is $\frac{1}{i(k-1)}$. Since both vertices are isolated, they can be included in either order, so the probability that e is added in the next step is $\frac{2}{i(k-1)}$.

If an external edge is added, the size of the gap decreases by one. If an internal edge is added, the gap is split in two gaps containing $i-2$ isolated vertices altogether. The sizes of the individual gaps are $j-1$ and $i-j-1$, respectively, where $1 \leq j \leq i$, and every choice of j has the same probability. Hence the following recursion holds for $i \geq 1$:

$$f_{k,i} = \frac{2f_{k,i-1}}{i(k-1)} + \sum_{j=1}^{i} \frac{2f_{k,j-1}f_{k,i-j-1}}{i(k-1)}. \tag{4.4}$$

If $i = 0$, then G_i consists of a path of length $k-1$, and only one edge is needed to complete C. Suppose e is the missing edge. The endpoints of e are the only ones of degree 1, so certainly one of them is chosen as the first vertex. The probability that the other is chosen as the second vertex is $\frac{1}{k-1}$; thus $f_{k,0} = \frac{1}{k-1}$.

For convenience we would like to have variables which are independent of k, so we let $g_i = (k-1)^{i+1}f_{k,i}$, where k is arbitrarily chosen. It is easy to see from (4.4) that g_i is independent of k, and we get the recursive formula

$$g_i = \frac{2}{i}\left(g_{i-1} + \sum_{j=1}^{i} g_{j-1}g_{i-j-1}\right) \tag{4.5}$$

and $g_0 = 1$. Let

$$g(z) = \sum_i g_i z^i.$$

Multiplying (4.5) by iz^i and summing over i, we get

$$\sum_i i g_i z^i \quad = \quad 2\sum_i g_{i-1}z^i + 2\sum_i \sum_{j=1}^{i} g_{j-1}g_{i-j-1}z^i,$$

which implies

$$g'(z) = 2g(z) + 2zg(z)^2.$$

Solving this differential equation, we get

$$g(z) = \frac{2}{1 - 2z + ce^{-2z}},$$

for a constant c. Since $g_0 = 1$, we must have $g(0) = 1$, so

$$g(z) = \frac{2}{1 - 2z + e^{-2z}}.$$

If we now go back to the beginning of the process, $G_{\min}(k, 0)$, the probability that the first edge added is an edge in C is $2/(k-1)$, and if this happens, the probability that the process becomes C at the end is $f_{k,k-2}$. Thus

$$\mathbb{P}\left[G_{\min}(k, k) = C\right] = \frac{2f_{k,k-2}}{k-1} = \frac{2g_{k-2}}{(k-1)^k}. \tag{4.6}$$

The Taylor expansion of $g(z)$ is

$$g(z) = 1 + 2z + 3z^2 + \frac{14}{3}z^3 + \frac{22}{3}z^4 + \frac{172}{15}z^5 + \frac{269}{15}z^6 + \dots.$$

Using the Taylor expansion and the equations (4.2) and (4.6), we can calculate the probability that $G_{\min}(k, k)$ is a cycle for small k, as found in Table 4.1.

k	probability	approximation
2	1	1.000
3	1/2	0.5000
4	2/9	0.2222
5	7/64	0.1094
6	176/3125	0.0563
7	43/1458	0.0295
8	12912/823543	0.0157
9	4419/524288	0.0084

Table 4.1: The probability that $G_{\min}(k, k)$ is a cycle.

Although little can be concluded about the asymptotic behaviour of this probability just by considering the initial values, it appears from Table 4.1 that the probability is roughly halved at every step. By considering the dominant singularity of $g(z)$, as explained in Section 2.1, we can show that this is not very far from the truth. The dominant singularity of $g(z)$ is $\frac{1+\alpha}{2}$, where $\alpha \approx 0.2785$ is the principal solution to $\alpha e^{\alpha+1} = 1$. Using Stirling's formula together with (4.2), we get

$$\mathbb{P}\left[G_{\min}(k, k) \text{ is a cycle}\right] \bowtie \left(\frac{2}{e(1+\alpha)}\right)^k,$$

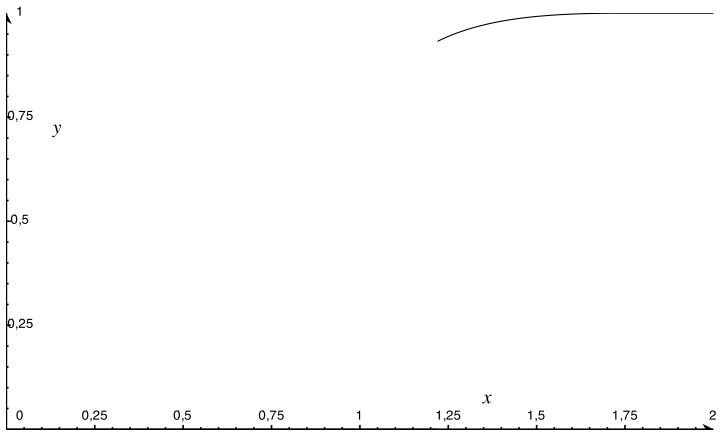

Figure 4.2: The probability that $G_{\min}(n, tn)$ is connected

where $\frac{2}{e(1+\alpha)} \approx 0.5755$.

Our main interest, however, is to use the generating function $g(z)$ to calculate the exact value of $\rho(t)$. If we multiply $g(z)$ by z and integrate, we get

$$\int zg(z)dz = -\frac{1}{2}\log\left(1 + (1-2z)e^{2z}\right) + C,$$

which we use to calculate that

$$\sum_{k\geq 0} \frac{g_k}{k+2} z^{k+2} = -\frac{1}{2}\log(1 + (1-2z)e^{2z}) + \frac{\log 2}{2}.$$

Then, according to (4.3) and (4.6),

$$\begin{aligned}
\lambda(t) &= \sum_{k=2}^{\infty} \frac{(k-1)^k}{2k} h(t)^k \frac{2g_{k-2}}{(k-1)^k} \\
&= \sum_{k=2}^{\infty} \frac{g_{k-2}}{k} h(t)^k = \sum_{k\geq 0} \frac{g_k}{k+2} h(t)^{k+2} \\
&= -\frac{1}{2}\log(1 + (1-2h(t))e^{2h(t)}) + \frac{\log 2}{2}.
\end{aligned}$$

The limit probability that $G_{\min}(n, tn)$ is connected is therefore

$$\rho(t) = e^{-\lambda(t)} = \sqrt{\frac{1 + (1-2h(t))e^{2h(t)}}{2}}$$

for $h_2 < t < h_3$. This function is shown in Figure 4.2. We have $h(h_2) = \frac{1}{2(1+\log 2)}$, so

$$\begin{aligned}\rho^+ &= \lim_{t\to h_2^+} \rho(t) = e^{-\lambda(h_2)} \\ &= \sqrt{\frac{1}{2}\left(1+\left(1-\frac{1}{1+\log 2}\right)e^{\frac{1}{1+\log 2}}\right)} \approx 0.9325.\end{aligned}$$

4.2 Main theorem

Having calculated the connectivity probability, we turn our attention to the main topic of this chapter, namely the phase transition of the minimum degree graph process. We recall from Section 1.4 that if $m = \frac{cn}{2}$ with $c < 1$, then $G(n,m)$ a.a.s. consists of small components, while if $m = \frac{cn}{2}$ with $c > 1$, then $G(n,m)$ a.a.s. contains a unique giant component of linear order. We will show that the minimum degree graph process undergoes a similar transformation, but while $G(n,m)$ needs only roughly $0.5n$ edges to produce the giant component, $G_{\min}(n,m)$ requires roughly $0.86n$ edges. Thus, while the mechanism employed to choose edges in $G_{\min}(n,m)$ makes the graph become connected much earlier than $G(n,m)$, the emergence of the giant component is slowed down. We will also show that the phase transition in $G_{\min}(n,m)$ happens as a double jump, similar to $G(n,m)$.

Theorem 4.2. *Let*

$$h_g = \log \frac{16\log 2 - 2}{3\log 2 - 1 + \log 2 \cdot \sqrt{27 - 16\log 2}} \approx 0.8607.$$

(i) If $t < h_g$, then a.a.s. every component in $G_{\min}(n,tn)$ has $O(\log n)$ vertices.

(ii) If $t = h_g$ and $\omega(n) \to \infty$, then $G_{\min}(n,tn)$ a.a.s. contains no component of order greater than $n^{2/3}\omega(n)$ and at least one component of order greater than $n^{2/3}/\omega(n)$.

(iii) If $t > h_g$, then a.a.s. the largest component in $G_{\min}(n,tn)$ has $s(t)n + o(n)$ vertices, where $s(t)$ is a function depending only on t, and every other component, if any, has $O(\log n)$ vertices.

We will prove Theorem 4.2 by approximating the graph process by a branching process: starting with a vertex v, we expose the component containing v, starting first with the vertices closest to v and then continuing as new vertices are added to the component.

One problem we have to circumvent is that given two incident edges, the event that one of them is in $G_{\min}(n,m)$ is not independent of the other being in the graph. We will overcome this problem by partitioning the edges into two classes, *red* and *blue*, according to the time they were added to the process and partitioning the vertices into *light* and *heavy* vertices. These concepts will be defined below.

We consider the process advancing in two separate phases. The first phase is when the minimum degree is 0, and the other is when the minimum degree is 1. We will represent the phases by colouring the edges in the following way: when an edge is added, we colour it *red* if the minimum degree of the graph (before the addition of the edge) is 0 and *blue* if the minimum degree is 1. Other edges are uncoloured, but we will only consider the stages of the process where all the edges are a.a.s. either red or blue, namely when $t < h_2$. We let the *red phase* be the part of the process where the graph still contains isolated vertices. The *blue phase* is the phase where the minimum degree is 1.

In the red phase the graph process behaves just like the random forest we considered in Chapter 3. Thus we can apply the results from that chapter to the red phase. In the blue phase the graph consists of a static red subgraph and an evolving blue subgraph. Since the first vertex chosen at every step in the blue phase must have degree 1, it must necessarily be incident to one red edge and no blue edges. Thus, no blue cycle can be formed, and the graph is in fact a union of a red and a blue forest.

In the blue phase every vertex is incident to a red edge, but not necessarily to a blue edge. Let v be a vertex. We let $C_{\text{red}}(v)$ be the maximal red tree containing v. If v is incident to at least one blue edge, we let $C_{\text{blue}}(v)$ be the maximal blue tree containing v. Otherwise, we let $C_{\text{blue}}(v) = \{v\}$ and consider it to be a blue tree of order 1. Thus every vertex is part of a red tree of order at least 2 and a blue tree of order at least 1. We also define the *red* and the *blue* degree of a vertex v to be the number of respectively red and blue edges incident to v; these are denoted $\deg_{\text{red}}(v)$ and $\deg_{\text{blue}}(v)$ respectively. We call v a *light* vertex if it has red degree 1 and *heavy* otherwise.

The crucial observation is that if we are given the information about whether v is light or heavy, then the order of $C_{\text{red}}(v)$ and the order of $C_{\text{blue}}(v)$ are two random variables which are essentially independent of each other. The reason for this is that when we add a new blue edge, which vertices we choose does not depend on the order of the components which they are part of, but only on the degrees of the vertices.

In the branching process we build up the component containing v by alternatingly adding red and blue trees. As t grows, the branching process will produce larger components, and the value h_g given in Theorem 4.2 cor-

responds to the critical point of the branching process: when $t \leq h_g$, the branching process dies out after a finite number of steps with probability 1, but when $t > h_g$, it continues forever with probability strictly greater than 0. We will show in Section 4.6 that this corresponds to the existence of a giant component in $G_{\min}(n, tn)$.

The rest of the chapter is organised as follows. In Sections 4.3 and 4.4 we discuss the order and structure of the red and blue trees respectively. Section 4.5 deals with the branching process. In Section 4.6 we use the branching process to prove the occurrence of the phase transition and the existence of the giant component in $G_{\min}(n, m)$ (part (i) and (iii) of Theorem 4.2), and in Section 4.7 we consider the critical moment (part (ii) of Theorem 4.2). Section 4.8 contains some remarks regarding the evolution of the minimum degree graph process.

4.3 The red phase

In Section 3.3 we showed how many red trees there are in $G_{\min}(n, m)$ at the end of the red phase. We only state the result here as it was given in Theorem 3.6.

Lemma 4.3. *When the red phase is finished, the number of red trees with exactly k vertices is a.a.s.*

$$\frac{k-1}{k2^k}n + o(n). \tag{4.7}$$

This lemma implies that for every vertex v,

$$\mathbb{P}\left[|C_{\text{red}}(v)| = k\right] = \frac{k-1}{2^k} + o(1). \tag{4.8}$$

In order to distinguish between light and heavy vertices, we have to distinguish between the red trees, not only by the number of vertices, but also according to the number of leaves. We recall from Section 3.2 that $e_{k,p}$ is the probability that a red tree of order k has exactly p leaves. The generating function for $E(u, v) = \sum_{k,p} e_{k,p} u^k v^p$ was given in Theorem 3.1.

When $t = h_1$, there are a.a.s. $n \log 2 + o(n)$ vertices of degree 1. Thus, when $t > h_1$, there are a.a.s. $n \log 2 + o(n)$ vertices incident to precisely one red edge (light vertices). From this and (4.8) it follows that

$$\mathbb{P}\left[C_{\text{red}}(v) \text{ is a } (k,p)\text{-tree} \middle| \deg_{\text{red}}(v) = 1\right] = \frac{p(k-1)}{k2^k \log 2} e_{k,p} + o(1) \tag{4.9}$$

and

$$\mathbb{P}\left[C_{\mathrm{red}}(v) \text{ is a } (k,p)\text{-tree} \middle| \deg_{\mathrm{red}}(v) > 1\right] = \frac{(k-p)(k-1)}{k2^k(1-\log 2)} e_{k,p} + o(1). \quad (4.10)$$

4.4 The blue phase

Now we will assume that we are somewhere in the blue phase; that is $h_1 < t < h_2$. Recall that a vertex is *light* if it is incident to precisely one red edge and *heavy* otherwise. Every nontrivial blue tree (i.e. a tree which is not a single vertex) begins as an edge and then possibly continues to grow one vertex at a time. When a nontrivial blue tree is first created, at most one of the two vertices in the tree can be heavy. Every subsequent vertex added to the tree must be light. Hence a blue tree cannot contain more than one heavy vertex. We say that a blue tree in $G_{\min}(n,m)$ is *simple* if every vertex in the tree is light, and *non-simple* otherwise. A non-simple tree must contain precisely one heavy vertex.

We will now determine how many simple and non-simple blue trees there are in $G_{\min}(n,tn)$. In order to simplify the formulas, we define

$$u = u(t) := 2e^{-t}.$$

Lemma 4.4. *The number of simple blue trees with exactly k vertices in $G_{\min}(n,tn)$ is a.a.s.*

$$\frac{1}{k}(1-u)^{k-1}(u + ku\log 2 - 1)n + o(n). \quad (4.11)$$

The number of non-simple blue trees with exactly k vertices in $G_{\min}(n,tn)$ is a.a.s.

$$(1-\log 2)u(1-u)^{k-1}n + o(n). \quad (4.12)$$

Proof. Let $S_k(n,m)$ be the number of simple blue trees with exactly k vertices, and $T_k(n,m)$ be the number of non-simple blue trees with exactly k vertices, in $G_{\min}(n,m)$. Lemma 3.5 can be adapted to blue trees as well as red, so we can assume that $k = O(\log n)$. We will use Theorem 2.5 to find functions $\sigma_k(t)$ and $\tau_k(t)$ for $k \geq 1$, such that a.a.s. $S_k(n,tn) = \sigma_k(t)n + o(n)$ and $T_k(n,tn) = \tau_k(t)n + o(n)$.

Consider a blue tree of order $k \geq 2$. The probability that this tree grows with one vertex when an edge is added to $G_{\min}(n,m)$ is the same as the probability that a red tree of order k grows with one vertex in the red phase.

Hence, for $k \geq 2$, (3.7) holds for blue trees as well as red, with $C_k(n,m)$ substituted with $S_k(n,m)$ and $T_k(n,m)$, respectively. Hence $\sigma_k(t)$ and $\tau_k(t)$ both satisfy the differential equation (3.8), and therefore also the recursive formula (3.12), with $\beta_k(t)$ substituted with $\sigma_k(t)$ and $\tau_k(t)$, respectively. Thus for $k \geq 2$,

$$\sigma_k(t) = (k-1)e^{-kt}\int \sigma_{k-1}(t)e^{kt}dt, \tag{4.13}$$

$$\tau_k(t) = (k-1)e^{-kt}\int \tau_{k-1}(t)e^{kt}dt. \tag{4.14}$$

The behaviour of the blue trees deviates from the red trees when $k = 1$. Every edge added causes a simple blue tree of order one to disappear. The expected amount of change of $S_1(n,m)$ and $T_1(n,m)$ are given by the equations

$$\mathbb{E}\left[S_1(n,m+1) - S_1(n,m)|G_{\min}(n,m)\right] = -1 - \frac{S_1(n,m)}{n} + o(1)$$

and

$$\mathbb{E}\left[T_1(n,m+1) - T_1(n,m)|G_{\min}(n,m)\right] = -\frac{T_1(n,m)}{n} + o(1).$$

Hence the differential equations

$$\frac{d}{dt}\sigma_1(t) = -1 - \sigma_1(t) \tag{4.15}$$

and

$$\frac{d}{dt}\tau_1(t) = -\tau_1(t) \tag{4.16}$$

are satisfied. As mentioned in the previous section, it was proved in [30] that when $t = h_1 = \log 2$, there are a.a.s. $n\log 2 + o(n)$ vertices of degree 1 in $G_{\min}(n, tn)$. Hence we have the boundary conditions $\sigma_1(\log 2) = \log 2$ and $\tau_1(\log 2) = 1 - \log 2$. Furthermore $\sigma_k(\log 2) = \tau_k(\log 2) = 0$ for $k \geq 2$. Solving the differential equations (4.15) and (4.16) and using the boundary conditions just given, we get

$$\begin{aligned}\sigma_1(t) &= 2(1+\log 2)e^{-t} - 1,\\ \tau_1(t) &= 2(1-\log 2)e^{-t}.\end{aligned}$$

Then, using the recursive formulas (4.13) and (4.14), it follows by induction that

$$\begin{aligned}\sigma_k(t) &= \frac{1}{k}(1-u)^{k-1}(u + ku\log 2 - 1),\\ \tau_k(t) &= (1-\log 2)u(1-u)^{k-1}.\end{aligned}$$

By Theorem 2.5, this implies the lemma. □

Since the number of vertices of red degree one when $t > h_1$ is a.a.s. $n \log 2 + o(n)$, Lemma 4.4 implies the following:

$$\mathbb{P}\left[|C_{\text{blue}}(v)| = k \middle| \deg_{\text{red}}(v) > 1\right] = u(1-u)^{k-1} + o(1), \tag{4.17}$$

$$\begin{aligned}&\mathbb{P}\left[C_{\text{blue}}(v) \text{ is simple and } |C(v)| = k \middle| \deg_{\text{red}}(v) = 1\right]\\ &= \frac{1}{\log 2}(1-u)^{k-1}(u + ku\log 2 - 1) + o(1),\end{aligned} \tag{4.18}$$

$$\begin{aligned}&\mathbb{P}\left[C_{\text{blue}}(v) \text{ is non-simple and } |C(v)| = k \middle| \deg_{\text{red}}(v) = 1\right]\\ &= \left(\frac{1}{\log 2} - 1\right)(k-1)u(1-u)^{k-1} + o(1).\end{aligned} \tag{4.19}$$

4.5 Branching process

In this section we explain how we build the components in $G_{\min}(n, m)$ using a multitype branching process. Multitype branching processes were introduced in Section 2.2.3.

In the branching process the vertices are of four different types. A vertex is either light or heavy, and it is either an *r-vertex* or a *b-vertex*. These will be defined below. Furthermore, we distinguish between *saturated* and *unsaturated* vertices.

Let $\mathcal{A}$ be the event that the number of red trees in $G_{\min}(n, tn)$ is given by (4.7), that the number of simple blue trees is given by (4.11), and that the number of non-simple blue trees is given by (4.12). According to Lemma 4.3 and Lemma 4.4, $\mathcal{A}$ holds with probability tending to 1. Thus, if we want to show that some event holds a.a.s., it is sufficient to show that it holds a.a.s. when conditioned on $\mathcal{A}$. From now on we therefore assume that the event $\mathcal{A}$ holds.

In the first step of the branching process a red tree is created. The order of the tree is given by the probability distribution (4.8). All the vertices created in this step are *unsaturated r-vertices*. If the tree has order k, we let p of the vertices be light and $k - p$ heavy, with probability $e_{k,p}$.

In subsequent steps the branching process evolves as follows. We choose an unsaturated vertex v at random. If v is an r-vertex, we create a blue tree incident to v. If v is heavy, then the order of the tree is chosen randomly with probabilities given by (4.17). If v is light, then the order of the tree, and whether it is simple or non-simple, is determined according to the probabilities in (4.18) and (4.19). All the vertices created are *unsaturated b-vertices*. Note that it is possible that the tree created has order 1; in this case no new

vertices are generated. If v is a heavy vertex, then all the newly generated vertices in the blue tree are light. If v is light and the blue tree is simple, all the new vertices are also light, and if the blue tree is non-simple, exactly one of the new vertices is heavy, while the others are light.

If, on the other hand, v is a b-vertex, we create a red tree incident to v. The probability that the red tree has order k and contains exactly p light vertices is given by the probability distribution (4.9) if v is a light vertex and (4.10) if v is heavy. All the newly created vertices are unsaturated r-vertices. After the new red or blue tree has been created, we end the step by marking v as *saturated*.

Thus a vertex is an r-vertex, if it was generated through the creation of a red tree, and a b-vertex if it was generated through the creation of a blue tree. As before, a vertex is *light* if it is incident to exactly one red edge, and *heavy* otherwise.

This branching process approximates the way we might proceed if we want to find all the vertices in the component in $G_{\min}(n, m)$ which contains a given vertex v. First we find the red tree containing v. Then at every vertex w of this red tree, we find the blue tree which contains w. As explained in Section 4.2, the order of the blue tree does not depend on the order of the red tree of which w is a part, only on the information about whether w is incident to one or more than one red edge. Then we continue exploring alternatingly red and blue trees, until we have found all the vertices in the component. Some care must be taken, because the graph process will generally contain cycles, which the branching process does not. We will consider this problem more closely in Sections 4.6 and 4.7.

We will now calculate the generating functions for the number of vertices created in one step of the branching process. By one step it is meant that from a given vertex we create a (red or blue) tree and then count the number of vertices which have been created. In subscripts we will often use the letters r, R, b and B. The letters r and b refer to light r- and b-vertices, whereas R and B refer to heavy r- and b-vertices.

Using (4.18) and (4.19) we get

$$\begin{aligned}
& f_r(x,y) \\
= & \sum_{i,j} \mathbb{P}\left[\text{a light } r\text{-vertex generates } i \text{ light and } j \text{ heavy } b\text{-vertices}\right] x^i y^j \\
= & \sum_k \Bigg(\mathbb{P}\left[C_{\text{blue}}(v) \text{ is simple and } |C(v)| = k+1 \middle| \deg_{\text{red}}(v) = 1\right] x^k \\
& + \mathbb{P}\left[C_{\text{blue}}(v) \text{ is non-simple and } |C(v)| = k+2 \middle| \deg_{\text{red}}(v) = 1\right] x^k y \Bigg) \\
= & \sum_k \Bigg(\frac{1}{\log 2}(1-u)^k (u + (k+1)u\log 2 - 1) x^k \\
& + (k+1)\left(1 - \frac{1}{\log 2}\right) u(1-u)^{k+1} x^k y \Bigg) \\
= & \frac{(y\log 2 - y + x)u^2 + ((1-\log 2)y + 1 + \log 2 - 2x)u - 1 + x}{(-xu + x - 1)^2 \log 2}.
\end{aligned}$$

By (4.17),

$$\begin{aligned}
& f_R(x,y) \\
= & \sum_{i,j} \mathbb{P}\left[\text{a heavy } r\text{-vertex generates } i \text{ light and } j \text{ heavy } b\text{-vertices}\right] x^i y^j \\
= & \sum_k \mathbb{P}\left[|C_{\text{blue}}(v)| = k+1 \middle| \deg_{\text{red}}(v) > 1\right] x^k \\
= & \sum_k u(1-u)^k x^k \\
= & \frac{u}{1-(1-u)x}.
\end{aligned}$$

In order to shorten the formulas and make them more readable, we set $\alpha = (x-2)(x-y)$, $\beta = (y-2)(y-x)$ and $\gamma = \frac{x}{y} e^{\frac{y-x}{2}}$. We will also in some

places write $z = \frac{y}{2}$. Using (3.2) and (4.9), we get

$$\begin{aligned}
& f_b(x,y) \\
= & \sum_{i,j} \mathbb{P}\left[\text{a light } b\text{-vertex generates } i \text{ light and } j \text{ heavy } r\text{-vertices}\right] x^i y^j \\
= & \sum_{k\geq 2}\sum_{p=2}^{k} \mathbb{P}\left[C_{\mathrm{red}}(v) \text{ is a } (k,p)\text{-tree} \middle| \deg_{\mathrm{red}}(v)=1\right] x^{p-1}y^{k-p} \\
= & \sum_{k\geq 2}\sum_{p\geq 2} \frac{p(k-1)}{k2^k \log 2} e_{k,p} x^{p-1} y^{k-p} \\
= & \frac{1}{x\log 2}\sum_{p\geq 2} p\left(\frac{x}{y}\right)^p \left(E_p(z) - \int \frac{E_p(z)}{z} dz\right) \\
= & \frac{1}{x\log 2}\sum_{p\geq 2} p\left(\frac{x}{y}\right)^p \left(\frac{p-1}{p^2} + \sum_{i=2}^{p} h_{p,i}(z)(i-1)(p-i+\frac{iy}{2}-1)\right) \\
= & \frac{1}{x\log 2}\left(\frac{x}{y-x} + \frac{-4\gamma+4\gamma^2+\alpha\gamma^2}{4(1-\gamma)^2} + \log\frac{y-x}{y(1-\gamma)}\right).
\end{aligned}$$

Finally we use (4.10) to calculate the last generating function:

$$\begin{aligned}
& f_B(x,y) \\
= & \sum_{i,j} \mathbb{P}\left[\text{a heavy } b\text{-vertex generates } i \text{ light and } j \text{ heavy } r\text{-vertices}\right] x^i y^j \\
= & \sum_{k\geq 2}\sum_{p=2}^{k} \mathbb{P}\left[C_{\mathrm{red}}(v) \text{ is a } (k,p)\text{-tree} \middle| \deg_{\mathrm{red}}(v)>1\right] x^{p}y^{k-p-1} \\
= & \sum_{k\geq 2}\sum_{p\geq 2} \frac{(k-p)(k-1)}{(1-\log 2)k2^k} e_{k,p} x^{p} y^{k-p-1} \\
= & \frac{1}{y(1-\log 2)}\sum_{p\geq 2} \left(\frac{x}{y}\right)^p \left(\frac{y}{2}E_p'(z) - (p+1)E_p(z) + p\int \frac{E_p(z)}{z} dz\right) \\
= & \frac{1}{y(1-\log 2)}\sum_{p\geq 2} \left(\frac{x}{y}\right)^p \Bigg(-\frac{p-1}{p} \\
& + \sum_{i=2}^{p} h_{p,i}(z) i(i-1)\left((p-i)\left(\frac{y}{2}-1\right) + \frac{iy^2}{4} - \frac{iy}{2} + 1\right)\Bigg) \\
= & \frac{1}{y(1-\log 2)}\left(\log\frac{y}{y-x} + \frac{x}{x-y} + \frac{\beta\gamma^2-4\gamma^2+4\gamma}{4(1-\gamma)^2} + \log(1-\gamma)\right).
\end{aligned}$$

We let T be the transition matrix associated with the branching process.

It can be written as

$$T = \begin{bmatrix} 0 & 0 & \phi_{1,3}(t) & \phi_{1,4}(t) \\ 0 & 0 & \phi_{2,3}(t) & \phi_{2,4}(t) \\ \psi_{3,1} & \psi_{3,2} & 0 & 0 \\ \psi_{4,1} & \psi_{4,2} & 0 & 0 \end{bmatrix}.$$

Here the first row and the first column correspond to light r-vertices, the second row and the second column to heavy r-vertices, the third row and the third column to light b-vertices, and the fourth row and the fourth column to heavy b-vertices. The entries denote the expected number of vertices we get of the required type when we start with a single vertex of the given type. The zeroes in the matrix are there because we will never go from an r-vertex to an r-vertex, or from a b-vertex to a b-vertex.

The ϕ's describe the expected number of vertices generated by the creation of a blue tree at an r-vertex, so they depend on t. The ψ's, on the other hand, describe how many vertices are generated by the creation of a red tree at a b-vertex and do not depend on t, since we are after the red phase in the graph process, and no more red edges are added.

The ϕ's can be calculated by differentiating the corresponding generating functions and evaluating them at $(x, y) = (1, 1)$. We let $\zeta = \frac{1}{\log 2}$.

$$\begin{aligned} \phi_{1,3}(t) &= \frac{\partial}{\partial x} f_r(1,1) = (\zeta - 2) - \frac{2(\zeta - 1)}{u} + \frac{\zeta}{u^2} \\ \phi_{1,4}(t) &= \frac{\partial}{\partial y} f_r(1,1) = (\zeta - 1)\frac{1-u}{u} \\ \phi_{2,3}(t) &= \frac{\partial}{\partial x} f_R(1,1) = \frac{1-u}{u} \\ \phi_{2,4}(t) &= \frac{\partial}{\partial y} f_R(1,1) = 0. \end{aligned}$$

The fact that $\phi_{2,4}(t) = 0$ is explained by the fact that a blue tree can contain at most one heavy vertex, so we will never go from a heavy r-vertex to a heavy b-vertex.

It is slightly more difficult to calculate the ψ's, since $x - y$ occurs in the denominator of the expressions of $f_b(x, y)$ and $f_B(x, y)$, so we cannot simply evaluate them at $(1, 1)$. The limits exist, however, so we can calculate the ψ's by taking the limits at $(1, 1)$:

$$\begin{aligned}
\psi_{3,1} &= \lim_{(x,y)\to(1,1)} \frac{\partial}{\partial x} f_b(x,y) = \frac{27 - 16\log 2}{16\log 2} \\
\psi_{3,2} &= \lim_{(x,y)\to(1,1)} \frac{\partial}{\partial y} f_b(x,y) = \frac{13}{16\log 2} \\
\psi_{4,1} &= \lim_{(x,y)\to(1,1)} \frac{\partial}{\partial x} f_B(x,y) = \frac{13}{16(1-\log 2)} \\
\psi_{4,2} &= \lim_{(x,y)\to(1,1)} \frac{\partial}{\partial y} f_B(x,y) = \frac{16\log 2 - 5}{16(1-\log 2)}.
\end{aligned}$$

Thus we have expressions for all the entries in T. Since T is not strictly positive, we cannot apply Theorem 2.4 to it directly. Instead we consider the 2×2-submatrix A of T^2 consisting of the first and second rows and columns. This matrix is the transition matrix when we consider the branching process in two steps at a time: from an r-vertex we first generate a blue tree, and then we generate a red tree for each of the new vertices we get. We have

$$A = \begin{bmatrix} \phi_{1,3}(t)\psi_{3,1} + \phi_{1,4}(t)\psi_{4,1} & \phi_{1,3}(t)\psi_{3,2} + \phi_{1,4}(t)\psi_{4,2} \\ \phi_{2,3}(t)\psi_{3,1} & \phi_{2,3}(t)\psi_{3,2} \end{bmatrix}.$$

Let $\lambda_1(t)$ be the largest eigenvalue of A. Then $\lambda_1(t)$ is an increasing and continuous function of t, and we define h_g to be the value for which $\lambda_1(h_g) = 1$. According to Theorem 2.4, the branching process dies out with probability 1 when $t \le h_g$, and it continues forever with positive probability when $t > h_g$.

We write $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. When all the elements of A are positive, the largest eigenvalue is

$$\lambda_1(t) = \frac{1}{2}\left(a + d + \sqrt{(a-d)^2 + 4bc}\right).$$

Setting $\lambda_1(t) = 1$ gives

$$a + d - ad + bc - 1 = 0.$$

Substituting in the values of a, b, c and d, we end up with a quadratic equation in u,

$$r_2 u^2 + r_1 u + r_0 = 0, \tag{4.20}$$

where

$$\begin{aligned} r_0 &= \frac{\zeta^2}{2} + \zeta - 1, \\ r_1 &= -\zeta^2 + 3\zeta, \\ r_2 &= \frac{\zeta^2}{2} - 4\zeta. \end{aligned}$$

The solutions of (4.20) are

$$u_\pm = \frac{3\zeta - \zeta^2 \pm \sqrt{27\zeta^2 - 16\zeta}}{8\zeta - \zeta^2}.$$

We must have $u > 0$, so the solution relevant to us is

$$u_+ = \frac{3\zeta - \zeta^2 + \sqrt{27\zeta^2 - 16\zeta}}{8\zeta - \zeta^2} = \frac{3\log 2 - 1 + \log 2 \cdot \sqrt{27 - 16\log 2}}{8\log 2 - 1}.$$

Since $u = 2e^{-t}$, we get

$$h_g = \log \frac{2}{u_+} = \log \frac{16\log 2 - 2}{3\log 2 - 1 + \log 2 \cdot \sqrt{27 - 16\log 2}}.$$

Thus we have located the critical moment of the branching process, when the largest eigenvalue of the transition matrix becomes 1.

We can also use the generating functions to express the extinction probability $P(t)$ of the branching process. When $t \leq h_g$, we have $P(t) = 1$. Suppose now that $t > h_g$. We let $g_r(x, y)$ and $g_R(x, y)$ be the generating functions for the number of light and heavy r-vertices generated in two steps (first generating a blue tree and then red trees afterwards) starting with one light r-vertex and one heavy r-vertex, respectively. We have

$$\begin{aligned} g_r(x, y) &= f_r(f_b(x, y), f_B(x, y)), \\ g_R(x, y) &= f_R(f_b(x, y), f_B(x, y)). \end{aligned}$$

By Theorem 2.4 there is a unique solution to the equations

$$\begin{aligned} y_r &= g_r(y_r, y_R), \\ y_R &= g_R(y_r, y_R), \end{aligned}$$

satisfying $0 \leq y_r, y_R < 1$. Then y_r is the extinction probability of the branching process starting with a single light r-vertex, and y_R is the extinction probability when starting with a single heavy r-vertiex.

The very first step in the branching process consists of creating a red tree. We call this tree $C_1(v)$. The extinction probability of the branching process is therefore

$$\begin{aligned} P(t) &= \sum_{k,p} \mathbb{P}\left[C_1(v) \text{ is a } (k,p)\text{-tree}\right] y_r^p y_R^{k-p} \\ &= \sum_{k,p} \frac{k-1}{2^k} e_{k,p} y_r^p y_R^{k-p} = \frac{y_r y_R (y_r - y_R)^2 e^{\frac{y_R - y_r}{2}}}{4\left(y_r e^{\frac{y_R - y_r}{2}} - y_R\right)}. \end{aligned}$$

4.6 The phase transition

In this section we prove parts (i) and (iii) of Theorem 4.2, using the branching process defined in the previous section and the following lemma.

Lemma 4.5. *Let $X_1, X_2, \ldots$ be nonnegative, integral, mutually independent, identically distributed random variables with mean μ, and let $X = X^{(r)} = \sum_{i=1}^{r} X_i$. For any constants $\delta > 0$ and $a > 0$, there is a constant $c > 0$ such that if $r \geq c \log n$, then*

$$\mathbb{P}[X^{(r)} \geq r(\mu + \delta)] = o(n^{-a}) \tag{4.21}$$

and

$$\mathbb{P}[X^{(r)} \leq r(\mu - \delta)] = o(n^{-a}). \tag{4.22}$$

Proof. For any $u > 0$, by Markov's inequality

$$\begin{aligned} \mathbb{P}[X \geq r(\mu+\delta)] &= \mathbb{P}[e^{uX} \geq e^{ur(\mu+\delta)}] \\ &\leq \mathbb{E}\left[e^{uX}\right] e^{-ur(\mu+\delta)} \\ &= \left(\mathbb{E}\left[e^{uX_i}\right] e^{-u(\mu+\delta)}\right)^r. \end{aligned}$$

Let $f(u) = \mathbb{E}\left[e^{uX_i}\right] e^{-u(\mu+\delta)}$ and $g(u) = \mathbb{E}\left[e^{uX_i}\right]$. To prove (4.21), it is sufficient to show that there is a $u > 0$ such that $f(u) < 1$. Since $f(0) = 1$, we only have to show that the derivative of $f(u)$ at $u = 0$ is negative.

Differentiating $g(u)$, we get

$$g'(u) = \sum_{k \geq 0} k e^{uk} \mathbb{P}\left[X_i = k\right],$$

so

$$g'(0) = \sum_{k \geq 0} k \mathbb{P}\left[X_i = k\right] = \mu$$

and

$$f'(0) \;=\; g'(0) - g(0)(\mu+\delta) = \mu - \mu - \delta = -\delta < 0.$$

To prove (4.22), we fix an m and define random variables Y_i such that

$$Y_i = \begin{cases} m - X_i & \text{if } X_i \le m, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mu' = \mathbb{E}[Y_i]$. For every $\eta > 0$ we can choose m large enough so that $\sum_{k=0}^{m} k\mathbb{P}[X_i = k] \ge \mu - \eta$. Hence we can choose m so large that

$$\begin{aligned} \mu' = \mathbb{E}[Y_i] &= \sum_{k\ge 0} k\mathbb{P}[Y_i = k] = \sum_{k=0}^{m} (m-k)\mathbb{P}[X_i = k] \\ &= m\sum_{k=0}^{m} \mathbb{P}[X_i = k] - \sum_{k=0}^{m} k\mathbb{P}[X_i = k] \\ &\le m - \mu + \eta. \end{aligned}$$

Let $Y = Y^{(r)} = \sum_{i=1}^{r} Y_i$, and note that $Y \ge rm - X$. We set $\eta = \delta/2$. Then

$$\begin{aligned} \mathbb{P}[X \le r(\mu-\delta)] &= \mathbb{P}[rm - X \ge rm - r(\mu-\delta)] \\ &\le \mathbb{P}[Y \ge r(m - \mu + \delta)] \\ &\le \mathbb{P}[Y \ge r(\mu' + \delta/2)], \end{aligned}$$

by choosing m sufficiently large. By (4.21) this probability can be well enough bounded, by choosing a large enough c. □

Proof of Theorem 4.2 (i) and (iii). We first prove part (i). Assume that $t < h_g$. Let v be a vertex in $G_{\min}(n, tn)$, and let $C(v)$ be the component containing v. We will bound the probability that $C(v)$ has more than $O(\log n)$ vertices by using the branching process of the previous section.

Since every vertex in $G_{\min}(n, tn)$ is incident to at least one red edge, every b-vertex in the branching process gives rise to at least one r-vertex. The total number of vertices is therefore at most twice the number of r-vertices, so it is sufficient to count the number of r-vertices.

Let $A = A(t)$ be the 2×2 transition matrix defined in the previous section, and let $\lambda_1 = \lambda_1(t)$ be the largest eigenvalue of A. Since $t < h_g$, we have $\lambda_1 = 1 - \varepsilon$ for some $\varepsilon > 0$. For a fixed k, we let $[v_1 v_2]$ be the left eigenvector of A corresponding to the eigenvalue λ_1, such that $v_1 + v_2 = k$.

Imagine that we start with u_1 light r-vertices and u_2 heavy r-vertices. From each of these vertices we first generate a blue tree, and from each of

the new vertices we then generate a red tree. Let V_1 be the number of newly generated light r-vertices, and V_2 be the same for heavy r-vertices. We define $P(x_1, x_2) = \mathbb{P}[V_1 \geq v_1 \text{ or } V_2 \geq v_2 | u_1 = x_1, u_2 = x_2]$.

The probability that $C(v)$ has at least $2k$ vertices is bounded from above by the probability that the branching process generates at least k r-vertices. This probability is again bounded from above by

$$\max_{x_1 \leq v_1, x_2 \leq v_2} P(x_1, x_2) = P(v_1, v_2).$$

Set $u_1 = v_1$ and $u_2 = v_2$. Since $[v_1 v_2]$ is an eigenvector, $\mathbb{E}[V_1] = (1-\varepsilon)v_1$ and $\mathbb{E}[V_2] = (1-\varepsilon)v_2$.

Let $W_{r\to r}$ be the random variable denoting the number of light r-vertices generated from one light r-vertex. Similarly define $W_{r\to R}$, $W_{R\to r}$ and $W_{R\to R}$. Then $V_1 = \sum_{i=1}^{v_1} W_{r\to r,i} + \sum_{i=1}^{v_2} W_{R\to r,i}$ and $V_2 = \sum_{i=1}^{v_1} W_{r\to R,i} + \sum_{i=1}^{v_2} W_{R\to R,i}$.

Thus both V_1 and V_2 are the sum of two random variables, each of which is a sum of a number of nonnegative, integral, mutually independent and identically distributed random variables. Hence we can use Lemma 4.5 to show that if we set $k = c \log n$, we can always choose c so large that

$$\mathbb{P}[V_i \geq v_i] = o(n^{-1})$$

for $i = 1, 2$, and so $P(v_1, v_2) = o(n^{-1})$.

It follows that the expected number of vertices in components of order greater than $2k$ is $n \cdot o(n^{-1}) = o(1)$, which finishes the proof of part (i).

Now we prove part (iii). Assume that $t > h_g$. In the branching process we distinguish between *saturated* and *unsaturated* vertices as in Section 4.5. Saturated vertices are those from which we have already added a red or a blue tree, while unsaturated vertices are those which have been generated, but from which we have not yet generated a new tree.

One problem appears when we try to use the branching process to model the graph process: in the branching process newly generated vertices are always distinct from vertices generated earlier. In the graph process it may happen that a newly generated vertex is the same as one of the unsaturated vertices. If the number of vertices we already have in the component is k, then the probability that a given newly generated vertex is one of the unsaturated vertices is bounded above by $\frac{k}{n}$. As long as $k \ll n$, for any $\varepsilon' > 0$ this value is smaller than ε' for large n. We therefore introduce a modified branching process: whenever a new vertex is generated, we discard this vertex with probability $\varepsilon' > 0$. Since $t > h_g$ and the eigenvalue function is continuous, we can always find an ε' small enough that the largest eigenvalue remains strictly greater than one. We will from now on assume that we are using this

modified branching process. The largest eigenvalue of the transition matrix of the modified branching process is $\lambda_1 = 1 + \varepsilon$, with $\varepsilon > 0$.

We will now prove that there are constants $c > 0$ and $\delta > 0$, such that for any η with $0 < \eta < \frac{1}{2}$, the following is true. Let $k_- = c \log n$ and $k_+ = n^{1-\eta}$. The probability that there is a vertex v such that for some k with $k_- \leq k \leq k_+$, the branching process starting with v has fewer than δk unsaturated vertices after k steps, given that the branching process has not died out before k_- steps, is $o(1)$.

To prove this, we fix k with $k_- \leq k \leq k_+$ and assume that the branching process has not died out after k_- steps. The expected number of vertices generated from k vertices is $k(1 + \varepsilon)$. Let δ be a constant with $0 < \delta < \varepsilon$. Lemma 4.5 implies that the constant c, which k_- depends on, can be chosen so large that the probability that the branching process after k steps (where $k \geq k_-$) has fewer than $k(1 + \delta)$ vertices is $o\left(n^{-2}\right)$. At every step one unsaturated vertex becomes saturated, so the number of unsaturated vertices is then at least $k(1+\delta)-k = \delta k$. The probability that for some vertex v and some k with $k_- \leq k \leq k_+$ the number of unsaturated vertices is less than δk is then bounded from above by

$$n \sum_{k=k_-}^{k_+} o\left(n^{-2}\right) = o(1).$$

In other words, there is a.a.s. no component with between k_- and k_+ vertices.

We now condition on the events that there is no component with between k_- and k_+ vertices and that whenever we have exposed k vertices of a component with $k_- \leq k \leq k_+$, there are δk unsaturated vertices. We want to show that there is a.a.s. no more than one component with more than $k_+ = n^{1-\eta}$ vertices. Let v and w be two vertices belonging to components with at least k_+ vertices. We run the branching processes starting with v and w. By assumption, when we have reached k_+ vertices in each of the branching processes, each of them has δk_+ unsaturated vertices. By Lemma 4.5, the δk_+ unsaturated vertices in the component containing v generate at least $\delta' k_+$ vertices for some $\delta' > 0$. The probability that none of these vertices is one of the unsaturated vertices in w's component is bounded from above by

$$\left(\frac{n - \delta k_+}{n}\right)^{\delta' k_+} = \left(1 - \delta n^{-\eta}\right)^{\delta' n^{1-\eta}} \leq e^{-\delta\delta' n^{1-2\eta}} = o\left(n^{-2}\right).$$

So a.a.s., if v and w are in components with more than k_+ vertices, then v and w are in the same component.

We call a component *small* if it has less than k_- vertices and *large* if it has more than k_+ vertices. We let $\rho(n, t)$ be the probability that a vertex v

in $G_{\min}(n, tn)$ is in a small component. In Section 4.5 we proved that when $t > h_g$ the probability that the branching process dies out is $P(t)$, with $0 < P(t) < 1$. Thus $\rho(n,t)$ is bounded from below by $P(t) + o(1)$. We let $P_{\varepsilon'}(t)$ be the extinction probability of the modified branching process defined earlier in the proof, where a newly created vertex is discarded with probability ε'. As long as $\varepsilon' > 0$, $P_{\varepsilon'}(t)$ is an upper bound for the probability that $C(v)$ is small. As $\varepsilon' \to 0$, $P_{\varepsilon'}(t)$ converges to $P(t)$, so $\rho(n,t) \to P(t)$ as $n \to \infty$.

Letting Y be the number of vertices in small components, we get $\mathbb{E}[Y] = (P(t) + o(1))\, n$. Moreover,

$$\mathbb{E}[Y(Y-1)] \le n\rho(n,t)(k_- + n\rho(n - O(k_-), t) = (1 + o(1))\mathbb{E}[Y]^2,$$

so by Chebyshev's inequality, a.a.s. $Y = (P(t) + o(1))\, n$. We conclude that there is a.a.s. a unique component with more than $O(\log n)$ vertices, and this component has a.a.s. $(s(t) + o(1))n$ vertices, where $s(t) = 1 - P(t)$. □

4.7 The critical phase

In this section we prove part (ii) of Theorem 4.2, the case $t = h_g$; that is when $\lambda_1(t) = 1$. This is called the *critical phase*. It turns out that the largest component in this case has order roughly $n^{2/3}$. Thus a double jump occurs in the minimum degree graph process, as in $G(n,p)$.

It will be convenient to consider the branching process in Section 4.5 as a single-type, rather than a multitype, branching process. We still distinguish between light and heavy vertices and between r-vertices and b-vertices, but we only count one of the types, say the light r-vertices. A step in the branching process consists of taking a light r-vertex and generating a blue tree incident to it, and then generating a red tree incident to each of the newly generated b-vertices. However, instead of stopping here as we did in the previous section, if there are now heavy r-vertices, we continue generating trees from these vertices, until there are only light r-vertices.

We let p_i be the probability that one light r-vertex generates precisely i new light r-vertices in this process. We let $p(z) = \sum_{i \ge 0} p_i z^i$ be the corresponding generating function. Since we are at the critical moment of the branching process, when $\lambda_1(t) = 1$, the process dies out with probability 1 according to Theorem 2.4. Hence the number of vertices created throughout the process is finite with probability 1, and so $p(1) = 1$.

We let $h(z)$ be the generating function for the number of light r-vertices

that are created when starting with one *heavy* r-vertex. Then

$$\begin{aligned} p(z) &= g_r(z, h(z)), \\ h(z) &= g_R(z, h(z)), \end{aligned}$$

where the functions g_r and g_R were defined in Section 4.5. Now we let q_i be the probability that the branching process starting from one light r-vertex dies out after having produced precisely i light r-vertices, including the vertex we started with. Then

$$q(z) = zp(q(z)),$$

by the standard argument used in Section 2.2 and in the previous section.

In Theorem 2.1, take $y(z) = q(z)$ and $\phi(z) = p(z)$. The expected number of light r-vertices produced by a light r-vertex is given by $p'(1)$. The value $p'(1)$ is in fact an increasing, continuous function of t. If $p'(1) > 1$, the process continues forever with positive probability. Since this is not the case when $t = h_g$, we must have $p'(1) \leq 1$. But for any $t > h_g$, the process does continue forever with positive probability, so in this case $p'(1) > 1$. Because of continuity, we must therefore have $p'(1) = 1$ when $t = h_g$.

Hence, by Theorem 2.1, the dominant singularity of $q(z)$ is $\rho = 1$. To show that (2.1) holds for the coefficients of $q(z)$, we only have to show that $p''(1) \neq 0$. This holds, since

$$p''(1) = \sum_{i \geq 2} i(i-1)p_i > 0. \tag{4.23}$$

Thus from (2.1) we obtain

$$q_i \sim ci^{-\frac{3}{2}}.$$

Given a vertex v, we can build the component $C(v)$, as explained earlier, by starting with the red tree containing v, and then adding blue and red trees alternatingly. The vertices in $C(v)$ are then labelled light and heavy, and r-vertices and b-vertices. The partition into r-vertices and b-vertices depends, however, on which vertex v we start from.

The branching process is only an approximation of the process of exposing the components. In the branching process the probability that we choose a red or blue tree of order k remains fixed throughout, but in the graph, every time we choose a tree of some order, the number of such trees in the graph is reduced by one, so the probability that we choose such a tree again later on is changed slightly. Every time a tree of order k is chosen, the number of vertices in such trees is reduced by k. Therefore the number of vertices in large trees is reduced at a greater rate than the number of vertices in smaller

trees, so the expected growth of the component exposure process is bounded by the expected growth of the branching process.

Let $Q_k^{r\to r}$ be the probability that a branching process starting with one light r-vertex lasts until at least k light r-vertices have been created. Then

$$Q_k^{r\to r} = \sum_{i\geq k} q_i \sim \sum_{i\geq k} ci^{-\frac{3}{2}} \sim c\int_k^\infty x^{-\frac{3}{2}}dx = 2ck^{-\frac{1}{2}},$$

for a constant c. Similarly we let $Q_k^{R\to r}$ be the probability that a branching process starting with one heavy r-vertex lasts until at least k light r-vertices have been found. We can repeat the above calculations for the process starting with a heavy r-vertex instead of a light, and show that $Q_k^{R\to r} \asymp k^{-1/2}$.

Recall again that the first step of the exposure process is to find the maximal red tree $C_1(v)$ containing the vertex v. Suppose that $C_1(v)$ contains l vertices, $v_1, \dots, v_l$. Then we can split the exposure process into l distinct branches, each starting at v_i for some $i = 1, \dots, l$. The first step at each of the branches is to generate a blue tree, and then the process continues as usual. If the exposure process starting at v reveals at least k light vertices, it means that for at least one vertex v_i with $1 \leq i \leq l$, the branch of the exposure process starting at v_i exposes at least $\frac{k}{l}$ vertices. We showed above that this happens with probability $\Theta((k/l)^{-1/2})$ regardless of whether v_i is light or heavy. Thus, by (4.8), the probability that exposing the component $C(v)$ starting at the vertex v reveals at least k light r-vertices is bounded by

$$\sum_{l\geq 2} \frac{l(l-1)}{2^l}\left(\frac{k}{l}\right)^{-1/2} \leq k^{-1/2}\sum_{l\geq 2}\frac{l^{3/2}}{2^l} \asymp k^{-1/2}. \tag{4.24}$$

Let $X_k^{(r)}$ be the random variable denoting the number of vertices v in $G_{\min}(n, h_g n)$ such that exposing $C(v)$ starting at v reveals at least k light r-vertices. By (4.24) and Markov's inequality,

$$\mathbb{P}[X_k^{(r)} \geq k] \leq \frac{\mathbb{E}[X_k^{(r)}]}{k} \asymp \frac{nk^{-1/2}}{k} \asymp nk^{-3/2}.$$

Similarly we define $X_k^{(R)}$, $X_k^{(b)}$ and $X_k^{(B)}$, where for example $X_k^{(B)}$ denotes the number of vertices v in $G_{\min}(n, h_g n)$ such that exposing $C(v)$ starting at v reveals at least k heavy b-vertices. By similar calculations one can show that the probability that either of these random variables is larger than k is $\Theta(nk^{-3/2})$, although with different implicit constants in each case. Markov's inequality implies that the probability that at least one of the four random variables $X_k^{(r)}$, $X_k^{(R)}$, $X_k^{(b)}$ and $X_k^{(B)}$ has value at least k is also $\Theta(nk^{-3/2})$.

Now suppose that $G_{\min}(n, h_g n)$ contains a component C consisting of at least $4k$ vertices. Let C_r be the set of vertices v in C such that the branching process starting at v finds at least k light r-vertices. Analogously define the sets C_R, C_b and C_B. The exposure process starting at v labels every vertex in C as an r-vertex or a b-vertex. In addition every vertex in C is either light or heavy. Thus for every vertex v the exposure process causes a partition of the vertices of C into four classes. One of these classes must contain at least k vertices, so v is a member of at least one of the sets C_r, C_R, C_b and C_B. Since this holds for every vertex in C, at least one of the sets must contain at least k vertices. It follows that if there is a component in $G_{\min}(n, h_g n)$ with at least $4k$ vertices, then at least one of the random variables $X_k^{(r)}$, $X_k^{(R)}$, $X_k^{(b)}$ and $X_k^{(B)}$ has value at least k, the probability of which we have calculated to be $\Theta(nk^{-3/2})$.

Assume that $\omega(n) \to \infty$ and $k = n^{2/3}\omega(n)$. The probability that there is a component of order at least $4k$ is then at most

$$Cnk^{-3/2} = \frac{C}{\omega(n)^{3/2}} \to 0,$$

for some constant C.

Now we consider the lower bound on the order of the largest component. We will prove that for any function $\omega(n) \to \infty$, there is at least one component $C(v)$ which contains $n^{2/3}/\omega(n)$ light r-vertices.

Let $\rho'(n, k)$ be the probability that the exposure process starting at a given vertex v reveals at least k light r-vertices. As calculated above, the probability $\rho'(n, k)$ is bounded above by $Q_k = Q_k^{r \to r}$. To find a lower bound for $\rho'(n, k)$ we define a modified branching process, as in the previous section: whenever a new vertex is generated, we discard it with probability $n^{-1/3}\omega'(n)$, for some function $\omega'(n) \to \infty$.

When we build the component $C(v)$, a newly generated vertex w may coincide with an already generated, but unsaturated, vertex w'. In this case we cannot generate new trees from both w and from w' — we solve this by disregarding both the vertex w and the vertex w'. The probability that a vertex has to be disregarded in this manner is bounded above by $n^{2/3}\omega'(n)/n = n^{-1/3}\omega'(n)$, since we have already proved that there are no components of order larger than $n^{2/3}\omega'(n)$.

Furthermore, as mentioned above, the probability that we choose a red or blue tree of some order l changes slightly throughout the process. Let for instance r_l be the probability that a given vertex is in a red tree of order l, at the beginning of the exposure. Thus the graph contains $r_l n$ vertices in such trees. After we have exposed $cn^{2/3}$ vertices, the expected number of vertices

exposed which are in red trees of order l is $r'_l n^{2/3}$, for some constant r'_l. If we now choose a vertex at random from the non-exposed vertices, the probability that this vertex is in a red tree of order k is therefore

$$\frac{r_l n - r'_l n^{2/3}}{n - cn^{2/3}} = r_l + O(n^{-1/3}).$$

Similar calculations hold if we instead consider the probability that a red (resp. blue) tree of order l is generated from a light or heavy b-vertex (resp. r-vertex). Thus, the probability that the branching process in one step chooses a tree of "wrong" order is bounded by $n^{-1/3}\omega'(n)$.

By discarding a newly generated vertex with probability $\delta := n^{-1/3}\omega'(n)$, the order of the components generated by the modified branching process therefore gives us a lower bound on the order of the components in the graph process. According to Lemma 3.5 there are a.a.s. no red trees in $G_{\min}(n,m)$ with order larger than $O(\log n)$; the proof of the lemma can easily be modified to hold for blue trees as well. In the modified branching process we therefore assume that no vertex gets more than $a \log n$ children, for some large enough constant a.

Now let $\omega(n) \to \infty$ and $k = n^{2/3}/\omega(n)$, and choose $\omega'(n)$ such that $\omega'(n)^2 = o(\omega(n))$, but $\omega'(n) \to \infty$. If we let $\overline{p}(z)$ be the generating function for the number of light r-vertices which are generated in one step from one light r-vertex in the modified branching process, we have

$$\begin{aligned}\overline{p}(z) &= \sum_{i=0}^{\lfloor a \log n \rfloor} p_i((1-\delta)z + \delta)^i \\ &= p((1-\delta)z + \delta).\end{aligned}$$

Defining $\overline{q}(z)$ analogously to $q(z)$, we get $\overline{q}(z) = zp((1-\delta)\overline{q}(z) + \delta)$. Let $\overline{\tau}$ be such that $\overline{p}(\overline{\tau}) - \overline{\tau}\overline{p}'(\overline{\tau}) = 0$ and let $\overline{\rho} = \frac{\overline{\tau}}{\overline{p}(\overline{\tau})}$ be the dominant singularity of $\overline{q}(z)$, as in Theorem 2.1. We have $\overline{p}(1) = p(1) = 1$ and $\overline{p}'(1) = (1-\delta)p'(1) = 1 - \delta$. Let $\overline{\delta} = \overline{\tau} - 1$. Then, using Lemma 2.2,

$$\begin{aligned}0 &= \overline{p}(\overline{\tau}) - \overline{\tau}\overline{p}'(\overline{\tau}) \\ &= \overline{p}(1) + \overline{\delta}\overline{p}'(1) + \frac{1}{2}\overline{\delta}^2\overline{p}''(1) + O(\overline{\delta}^3) \\ &\quad -(1+\overline{\delta})\left(\overline{p}'(1) + \overline{\delta}p''(1) + \frac{1}{2}\overline{\delta}^2\overline{p}'''(1) + O(\overline{\delta}^3)\right) \\ &= \delta - \overline{\delta}\overline{p}''(1) - \overline{\delta}^2\frac{1}{2}\left(\overline{p}''(1) + \overline{p}'''(1)\right) + O(\overline{\delta}^3)).\end{aligned}$$

We have $\overline{p}''(1) = (1-\delta)^2 p''(1) \sim p''(1) > 0$ by (4.23), so we conclude that $\overline{\delta} \asymp \delta$. Using the equation

$$\log(1+x) = x - \frac{x^2}{2} + O(x^3),$$

we calculate

$$\begin{aligned}
\log\overline{\rho} &= \log\frac{\overline{\tau}}{\overline{p}(\overline{\tau})} = \log\overline{\tau} - \log\overline{p}(\overline{\tau}) \\
&= \log(1+\overline{\delta}) - \log\overline{p}(1+\overline{\delta}) \\
&= \overline{\delta} - \frac{1}{2}\overline{\delta}^2 + O(\overline{\delta}^3) - \log\left(\overline{p}(1) + \overline{\delta}\overline{p}'(1) + \frac{1}{2}\overline{\delta}^2\overline{p}''(1) + O(\overline{\delta}^3)\right) \\
&= \overline{\delta} - \frac{1}{2}\overline{\delta}^2 + O(\overline{\delta}^3) - \left(\overline{\delta}\overline{p}'(1) + \frac{1}{2}\overline{\delta}^2\overline{p}''(1) - \frac{1}{2}\overline{\delta}^2\overline{p}'(1)^2 + O(\overline{\delta}^3)\right) \\
&= \frac{1}{2}\overline{\delta}^2\overline{p}''(1) + O(\overline{\delta}^3).
\end{aligned}$$

Hence $\log\overline{\rho} = c'(n^{-1/3}\omega'(n))^2$ for a constant c'. It follows that

$$\overline{\rho}^{-k} = e^{-c'kn^{-2/3}\omega'(n)^2} = e^{-c'\frac{\omega'(n)^2}{\omega(n)}} = e^{-o(1)} = 1 - o(1).$$

The constant in (2.1) depends only on the derivatives of $p(z)$, so we can conclude that when $n \to \infty$, we get $[z^i]\overline{q}(z) \sim [z^i]q(z)$. Thus $\rho'(n,k)$ is asymptotically bounded below by Q_k. This gives us that $\rho'(n,k) \sim Q_k$ and $\mathbb{E}[X_k^{(r)}] \sim nQ_k$.

We now condition on the event that there is no component with more than $k' = n^{2/3}\omega'(n)$ vertices; we have already proved that this event holds a.a.s. We want to prove that there is at least one component with at least $k = n^{2/3}/\omega(n)$ vertices. We have

$$\mathbb{E}[X_k^{(r)}] \sim nQ_k = 2cnk^{-\frac{1}{2}} = 2cn^{2/3}\sqrt{\omega(n)}.$$

Let v be a vertex such that the exposure process starting at v reveals at least k light r-vertices. By assumption, $C(v)$ contains at most k' vertices. Hence, the expected number of pairs of vertices, v and w, such that both the exposure process starting at v, and the one starting at w, reveal at least k light r-vertices, is bounded from above by $k' + \mathbb{E}[X_k^{(r)}]$.

$$\begin{aligned}
\mathbb{E}[X_k(X_k^{(r)} - 1)] &\leq \mathbb{E}[X_k^{(r)}]\left(k' + \mathbb{E}[X_k^{(r)}]\right) \\
&\sim 2cn^{2/3}\sqrt{\omega(n)}(n^{2/3}\omega'(n) + 2cn^{2/3}\sqrt{\omega(n)}) \\
&= \mathbb{E}[X_k^{(r)}]^2(1+o(1)),
\end{aligned}$$

so by Chebyshev's inequality, we get that a.a.s. $X_k^{(r)} \sim \mathbb{E}[X_k^{(r)}]$. In particular there is a.a.s. a component in $G_{\min}(n, h_g n)$ of order at least k.

4.8 The evolution of the minimum degree graph process

Using the results in [30], [29] and Chapters 3 and 4, we have a fairly complete picture of many aspects of the evolution of the minimum degree graph process. Erdős and Rényi identified in [18] five phases through which the graph process $G(n,m)$ passes. When $m = o(n)$, the graph $G(n,m)$ is a.a.s. a forest. When $m \sim cn$ with $0 < c < 1/2$, $G(n,m)$ consists a.a.s. of trees and unicyclic components and has no component with more than $O(\log n)$ vertices. When $m \sim cn$ with $c > 1/2$, $G(n,m)$ contains a giant component, which increases in order as c increases. In the phase where $m \sim cn \log n$ with $c \leq 1/2$, the graph a.a.s. becomes connected, and when $m \gg n \log n$, the graph is a.a.s. connected, and the degrees of all vertices are a.a.s. asymptotically equal.

We can similarly identify five phases through which $G_{\min}(n,m)$ passes, which remind of the phases for $G(n,m)$. As usual $m = tn$. The first phase is when $t < h_1$, which we have called the red phase in this chapter. In this phase the graph is a forest, whose trees grow by one vertex at a time. The second phase is when $h_2 < t < h_g$. Here the trees no longer grow with just one vertex at a time, but merge with each other, allowing them to grow much quicker and form cycles. However, a.a.s. no component has more than $O(\log n)$ vertices in this phase. The third phase is when $h_g < t < h_2$. In this phase the giant component has a.a.s. been formed, and it grows steadily as t increases. When t reaches h_2, the giant component comprises $n - o(n)$ vertices a.a.s., and the only vertices, if any, which have not been swallowed by the giant form isolated cycles. The fourth phase is when $h_2 < t < h_3$. The isolated cycles which may still exist live in constant danger of being merged into the giant component. If the graph is not already connected, it may become so at any time. The fifth phase is when $t > h_3$. By now the graph is a.a.s. connected, and the minimum degree of the graph process grows steadily as t increases.

Theorem 4.2 states that a double jump occurs in $G_{\min}(n,m)$ as in $G(n,m)$. This implies that the phase transition happens much in the same way in both processes, even though it occurs at different points in time: the phase transition happens after $0.5n$ edges in $G(n,m)$, but after roughly $0.86n$ edges in $G_{\min}(n,m)$. That the phase transition happens later in $G_{\min}(n,m)$ is not surprising, because after only $0.5n$ edges, the graph still contains isolated vertices, so the components can only grow by one vertex at a time.

It may be interesting to compare $G_{\min}(n,m)$ to another random graph model which shares some of the properties of $G_{\min}(n,m)$. In Section 1.5 we presented random graphs with a given degree sequence. Theorem 1.3

by Molloy and Reed [44] states when the phase transition occurs in this model. In order to compare $G_{\min}(n,m)$ with this graph model, we proceed as follows. We first generate the graph $G_{\min}(n,m)$, and let $\mathbf{d}_n(m)$ be its degree sequence. Then we choose a random graph G_n with degree sequence $\mathbf{d}_n(m)$. Thus G_n has asymptotically the same degree distribution as $G_{\min}(n,m)$, but is generated differently. We want to determine how large m must be for this random graph to produce a giant component.

In Section 2.3 we defined $X_k(n,m)$ to be the number of vertices in $G_{\min}(n,m)$ with degree k, and we showed that there are functions $\alpha_0(t), \alpha_1(t), \ldots$ such that a.a.s. $X_k(n,m) = \alpha_k(t)n + o(n)$. We also defined the generating function $A(z,t) = \sum_k \alpha_k(t) z^k$ and showed that

$$A(z,t) = \left(\frac{e^t}{2}\right)^{z-1} (z - 1 + 2^z) - z \tag{4.25}$$

when $h_1 < t < h_2$. Let $Q(t) = \sum_{i \geq 1} i(i-2)\alpha_i(t)$. It follows from Theorem 1.3 that the phase transition occurs when $Q(t) = 0$: if $Q(t) < 0$, we are before the phase transition, and all components are small, and if $Q(t) > 0$, we are after the phase transition, and there is a giant component. The equation $Q(t) = 0$ is equivalent to

$$\frac{\partial^2}{\partial z^2} A(z,t)\bigg|_{z=1} = \frac{\partial}{\partial z} A(z,t)\bigg|_{z=1}. \tag{4.26}$$

By differentiating (4.25) and evaluating at $z = 1$, (4.26) becomes

$$2t - 2\log 2 + 2t^2 = 2t.$$

From this we conclude that the phase transition in this random graph model occurs when $t = \sqrt{\log 2} \approx 0.8326$, slightly earlier than in the minimum degree graph process.

The research in this chapter was helped by simulations done by Zedeňek Petrašek, which demonstrated that the phase transition occurs near h_g. Using different techniques, Janson and Łuczak also found the approximate location of the phase transition.

Chapter 5

Random graphs with a given degree sequence

The model of random graphs with a given degree sequence was presented in Section 1.5. We repeat the basic definitions, which to a large extent are based on [44]. Let A be an infinite set of natural numbers. An *asymptotic degree sequence* is a sequence of integer-valued functions $\mathcal{D} = \{d_0(n), d_1(n), \ldots\}$, where $d_i(n) : A \to \mathbb{N}_0$, such that $d_i(n) = 0$ for $i \geq n$, and $\sum_{i \geq 0} d_i(n) = n$. The set A plays no important role and is only included in the definition to allow for certain classes of graphs, such as regular graphs with odd degree, which place some restrictions on the number of vertices. In this chapter we will simply assume that $A = \mathbb{N}$. If $\mathcal{D}$ is an asymptotic degree sequence, we let $\mathcal{D}_n$ be the degree sequence $(a_1, a_2, \ldots, a_n)$, where $a_j \leq a_{j+1}$ for every $j = 1, \ldots, n-1$, and $\#\{j | a_j = i\} = d_i(n)$. Thus $d_i(n)$ is the number of vertices of degree i in a graph of order n. Let $\Omega_{\mathcal{D}_n}$ be the set of graphs on n vertices with degree sequence $\mathcal{D}_n$. An asymptotic degree sequence $\mathcal{D}$ is *feasible* if $\Omega_{\mathcal{D}_n} \neq \emptyset$ for all $n \geq 1$. We will henceforth assume that every asymptotic degree sequence we encounter is feasible.

If $\mathcal{D}$ is an asymptotic degree sequence, we let $G_n = G_n(\mathcal{D})$ be a random graph chosen uniformly at random from the set $\Omega_{\mathcal{D}_n}$. We say that G_n is a random graph with the given (asymptotic) degree sequence $\mathcal{D}$, and we are interested in the properties of G_n as n tends to infinity.

5.1 Main theorem

Molloy and Reed [44] showed that this random graph model exhibits a phase transition phenomenon, similarly to the standard random graph $G(n,p)$: there are degree sequences $\mathcal{D}$ such that the random graph $G_n(\mathcal{D})$ typically

consists of many small components, but such that by changing the degree sequence by a small amount to $\mathcal{D}'$, the random graph $G_n(\mathcal{D}')$ typically contains a large component of linear order. In order to discuss this result, we need some more terminology, introduced by Molloy and Reed.

For a feasible asymptotic degree sequence $\mathcal{D} = \{d_0(n), d_1(n), \ldots\}$, let $\lambda_i(n) = d_i(n)/n$. Recall from Section 1.5 that $\mathcal{D}$ is *smooth* if there are constants λ_i^* such that $\lim_{n\to\infty} \lambda_i(n) = \lambda_i^*$ for $i \geq 0$.

We can now introduce the following definition, which is similar to the definition in [44], although slightly stronger.

Definition 5.1. *An asymptotic degree sequence $\mathcal{D}$ is well-behaved if the following conditions are satisfied.*

1. *$\mathcal{D}$ is feasible and smooth.*
2. *Let $f(i)$ be a polynomial in i with degree at most 3. Then*
   (a) *$f(i)\lambda_i(n)$ tends uniformly to $f(i)\lambda_i^*$: that is, for all $\varepsilon > 0$, there exists N such that for all $n \geq N$ and for all $i \geq 0$*

$$|f(i)\lambda_i(n) - f(i)\lambda_i^*| < \varepsilon.$$

   (b)

$$L_f(\mathcal{D}) = \lim_{n\to\infty} \sum_{i\geq 1} f(i)\lambda_i(n)$$

   exists, and the sum approaches the limit uniformly: that is, for all $\varepsilon > 0$, there exist i^* and N such that for all $n \geq N$

$$\left| \sum_{i=1}^{i^*} f(i)\lambda_i(n) - L_f(\mathcal{D}) \right| < \varepsilon.$$

In Section 1.5.2 we defined the quantity

$$Q(\mathcal{D}) = \sum_{i\geq 1} i(i-2)\lambda_i^*.$$

Theorem 1.3, which was proved by Molloy and Reed [44], states that the phase transition occurs when $Q(\mathcal{D}) = 0$: if $Q(\mathcal{D}) < 0$, all components in G_n are a.a.s. "small", while if $Q(\mathcal{D}) > 0$, there is a.a.s. a unique component in G_n with $\Theta(n)$ vertices. The theorem does not cover the case that $Q(\mathcal{D}) = 0$, which is usually referred to as the *critical point*. In this chapter we consider this case in more detail. We will assume that $\lambda_1^* > 0$, so that there is a nonvanishing proportion of vertices of degree 1. If $Q(\mathcal{D}) = 0$ and $\mathcal{D}$ is well-behaved, it follows that there must be a $j > 2$ such that $\lambda_j^* > 0$ as well. The

structure of the random graphs with a degree sequence such that $Q(\mathcal{D}) = 0$ depends on how fast the quantity $\sum_{i\geq 1} i(i-2)\lambda_i(n)$ converges to 0. In order to study this, we define

$$Q_n(z) = \sum_{i\geq 1} i(i-2)\lambda_i(n)z^i.$$

If $\lambda_j^* > 0$ for some $j > 2$, then $Q_n''(z) = \sum_{i\geq 3} i^2(i-1)(i-2)\lambda_i(n)z^{i-2}$ is positive when $z > 0$, and hence the function $Q_n(z)$ is convex on the interval $[0, \infty)$. It therefore has at most two zeroes on this interval. Let τ_n be the largest value such that

$$Q_n(\tau_n) = 0. \tag{5.1}$$

If $Q(\mathcal{D}) = 0$, then $\lim_{n\to\infty} \tau_n = 1$.

We define the generating function

$$\Lambda_n(z) = \sum_{i\geq 0} \lambda_i(n)z^i.$$

Then $Q_n(z)$ can be written as

$$Q_n(z) = z^2\Lambda_n''(z) - z\Lambda_n'(z). \tag{5.2}$$

Note that since $\mathcal{D}$ is well-behaved, $\Lambda_n'(1)$, $\Lambda_n''(1)$ and $\Lambda_n'''(1)$ are all bounded as $n \to \infty$. We let $d = \Lambda'(1)$ be the average degree of G_n.

We will prove the following theorem, which is comparable to the results regarding the critical phase of $G(n, p)$ by Bollobás [8] and Łuczak [35] (see Theorem 1.2 in Section 1.4).

Theorem 5.2. *Assume that $\mathcal{D}$ is a well-behaved asymptotic degree sequence, such that for some $\varepsilon > 0$, $d_i(n) = 0$ whenever $i > n^{1/4-\varepsilon}$. Furthermore assume that $\lim_{n\to\infty} \tau_n = 1$ and $\lambda_1^* > 0$. Let*

$$\delta_n = 1 - \tau_n. \tag{5.3}$$

(i) If $\delta_n n^{1/3} \to -\infty$, then a.a.s. all components in $G_n(\mathcal{D})$ have $o(n^{2/3})$ vertices.

(ii) There is a constant c_1 such that if $\delta_n n^{1/3} \geq c_1 \log n$ then a.a.s. $G_n(\mathcal{D})$ has a single component with $\gg n^{2/3}$ vertices, while all other components have $o(n^{2/3})$ vertices.

Note that $Q(\mathcal{D}) = \lim_{n\to\infty} Q_n(1)$. Since $Q(\mathcal{D})$ is the quantity used in the statement of Theorem 1.3, it may be interesting to see what Theorem 5.2 says about the quantity $Q_n(1)$. By Lemma 2.2 and (5.3), we see that

$$Q_n(\tau_n) = Q_n(1) - \delta_n Q_n'(1) + O(\delta_n^2).$$

Equation (5.1) then implies that

$$Q_n(1) \sim \delta_n Q_n'(1), \tag{5.4}$$

where $Q_n'(1)$ converges to a positive constant when $n \to \infty$. Theorem 5.2 can therefore alternatively be formulated such that $Q_n(1)n^{1/3} \to -\infty$ and $Q_n(1)n^{1/3} \geq c_1' \log n$, for some constant $c_1' > 0$, respectively. However, the quantity δ_n occurs naturally in the proof, so it simplifies the notation to state the theorem in terms of δ_n.

In order to prove Theorem 5.2, we will model the components by a branching process, which we will introduce in Section 5.2. In Section 5.3 we study the behaviour of the branching process by examining the behaviour of some generating functions associated with it. The proof of Theorem 5.2 follows in Section 5.4.

In Section 1.5.1 we introduced the configuration model which is often used to generate random graphs with a given degree sequence. Let us recall the procedure of constructing a random configuration. Suppose $\mathcal{D}_n = \{a_1, \dots, a_n\}$. We let $v_1, \dots, v_n$ be vertices and L_n be a set consisting of a_i distinct copies of the vertex v_i for $i = 1, \dots, n$. These copies are called *half-edges*. A random configuration $\mathcal{C}_n$ consists of L_n together with a perfect matching $\mathcal{P}_n$ of L_n, chosen uniformly at random.

Given a configuration $\mathcal{C}_n$ on n vertices, we let $G_n^* = G_n^*(\mathcal{C}_n)$ be the multigraph obtained by identifying all copies of v_i with one another for every $i = 1, \dots, n$, and letting the pairs of the perfect matching in $\mathcal{C}_n$ become edges. The multigraph G_n^* is then a random multigraph with degree sequence $\mathcal{D}_n$; however, it is not chosen uniformly at random.

In order to generate a random *simple* graph with degree sequence $\mathcal{D}_n$, we repeat the above procedure until we obtain a graph G_n^* without loops and multiple edges. As explained in Section 1.5.1, every simple graph with degree sequence $\mathcal{D}_n$ has the same probability of being chosen, so this procedure produces a simple graph with degree sequence $\mathcal{D}_n$ uniformly at random.

Assume that $d_i(n) = 0$ for all $i > n^{1/4-\varepsilon}$ for some $\varepsilon > 0$, and that $\mathcal{D}_n$ otherwise satisfies the conditions of Theorem 5.2. As observed in [44], the main result of [42] implies that G_n is a simple graph with probability tending to $e^{-\nu(\mathcal{D}_n)}$, for some $\nu(\mathcal{D}_n) = O(n^{1/2-\varepsilon})$. If $Q(\mathcal{D})$ is constant, then $\nu(\mathcal{D}_n)$ tends to a constant. This implies the following lemma:

Lemma 5.3 ([44])**.** *If a random configuration $\mathcal{C}_n$ with a given degree sequence $\mathcal{D}_n$ meeting the conditions of Theorem 5.2 has a property $\mathcal{P}$ a.a.s., then a random graph with the same degree sequence has $\mathcal{P}$ a.a.s.*

5.2 Branching process

We use a branching process in order to study the size of the components in the underlying multigraph of the random configuration, which we defined in the previous section.

Consider the set L_n. In Section 1.5.1 it was explained that the perfect matching $\mathcal{P}_n$ can be constructed greedily by choosing the pairs one by one at random. The first half-edge in a pair can be chosen arbitrarily, while the second half-edge must be chosen uniformly at random from the set of available half-edges. Thus we start by choosing a single pair of half-edges uniformly at random, forming the first edge in $\mathcal{P}_n$. We want to determine the size of the component containing this edge, and we will do this by *exposing* the pairs of the perfect matching $\mathcal{P}_n$. Suppose that the pair we choose contains the two half-edges v_1 and v_2. Then we say that v_1 and v_2 are *exposed*, while all other half-edges are *unexposed*. A vertex v (consisting of one or more half-edges) is *unexposed* if none of its half-edges are exposed, *partially exposed* if some, but not all, of its half-edges are exposed, and *fully exposed* if all its half-edges have been exposed.

The process of exposing the component containing the pair v_1v_2 goes on as follows. At every step we choose an unexposed half-edge w_1, randomly or otherwise, in any partially exposed vertex, provided that such a vertex exists. Then we choose another half-edge, w_2, uniformly at random from all unexposed half-edges in L_n. Then we add the pair w_1w_2 to the matching and say that w_1 and w_2 are exposed.

When there is no partially exposed vertex left in the configuration, we stop the process. The component containing the edge v_1v_2 has then been fully exposed, and the set of exposed vertices forms a connected component in G_n^*.

We will model this process as a branching process, where the particles in the branching process are edges. An edge consists of two half-edges, which we will call the *upper* and *lower* half-edge. In the branching process an edge gets i edges as children, for $i = 0, 1, 2, \ldots$, with probability $\frac{(i+1)\lambda_{i+1}}{d}$, where $d = \sum i\lambda_i(n) = \Lambda'(1)$ is the average degree. This equals the probability that a randomly chosen half-edge is a part of a vertex of degree $i + 1$. We will interpret the branching process such that the lower half-edge of an edge, together with the upper half-edges of all its children, comprises one vertex in the random graph.

The branching process starts with a single edge, v_1v_2, which is special in that we consider both the half-edges v_1 and v_2 to be lower half-edges. Thus the branching process starts off with two branches which continue independently of each other. Figure 5.1 shows the situation after the first step of the

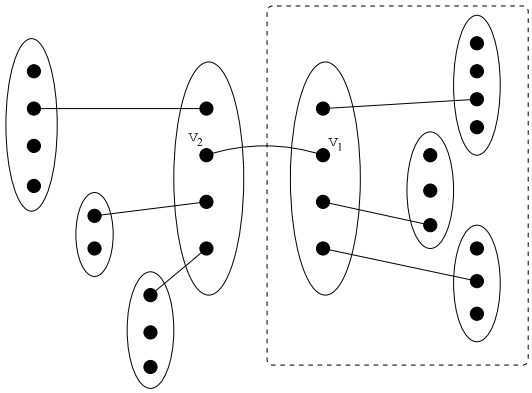

Figure 5.1: The first step of the branching process

branching process.

We let B_n be the random variable denoting the number of *vertices* produced in the branching process before it dies out. If the branching process does not die out, but continues forever, $B_n = \infty$.

Since the branching process starts with two independent branches, it will be convenient to consider the corresponding branching process, which starts with one edge v_1v_2, but only lets the branching process continue from one of the half-edges, say v_1, as in the boxed part of Figure 5.1. Let β_n be the random variable denoting the number of edges produced in this process, including the original edge. The total number of edges in the branching process is then $\beta_n^{(1)} + \beta_n^{(2)} - 1$, where $\beta_n^{(1)}$ and $\beta_n^{(2)}$ are independent random variables with the same distribution as β_n. The number of vertices produced in the process is then $B_n = \beta_n^{(1)} + \beta_n^{(2)}$.

We let $p_n(z)$ be the probability generating function for the number of children of an edge in the branching process. Then

$$p_n(z) = \sum_{i\geq 0} p_i(n) z^i = \sum_{i\geq 0} \frac{(i+1)\lambda_{i+1} z^i}{\sum_j j\lambda_j} = \frac{\Lambda_n'(z)}{\Lambda_n'(1)}. \tag{5.5}$$

The expected number of children of an edge is $p_n'(1)$. It then follows from Theorem 2.3 that if $p_n'(1) \leq 1$, then $\mathbb{P}[\beta_n < \infty] = \mathbb{P}[B_n < \infty] = 1$.

Let e be any edge in the branching process. Let $q_k(n)$ be the probability that the total number of descendant edges of e, including e itself, is exactly k. Let $q_n(z) = \sum q_k(n) z^k$ be the corresponding probability generating function. The probability that e has i children is $p_i(n)$, so

$$q_n(z) = z \sum_{i\geq 0} p_i(n) q_n(z)^i = z p_n(q_n(z)). \tag{5.6}$$

Since $B_n = \beta_n^{(1)} + \beta_n^{(2)}$, the probability generating function for B_n is $q_n(z)^2$.

There are two difficulties which are not taken account of by the branching process, compared with G_n. The first problem is that in the random graph G_n, or more precisely in the random configuration, a half-edge may choose to form an edge with a half-edge in a vertex which already contains exposed half-edges, which causes a cycle to be formed. We will show that this happens sufficiently seldom that the branching process is a good enough approximation. The second problem is that when some vertices are already partially exposed, the probability that a new half-edge is in a vertex of degree i generally deviates from $\frac{i\lambda_i}{d}$ in random configurations. However, we will see in Lemma 5.7 that this deviation is also small enough, and that the branching process is a good approximation.

5.3 Analysis of generating functions

In this section we will study the behaviour of the branching process. In particular we want to calculate the probability $\mathbb{P}[B_n \geq k]$ for various values of k. The event that $B_n \geq k$ can happen in two ways: either the branching process dies out after it has produced k vertices, or it continues forever. Hence

$$\mathbb{P}[B_n \geq k] = \mathbb{P}[k \leq B_n < \infty] + \mathbb{P}[B_n = \infty]. \tag{5.7}$$

In order to calculate this quantity, we use Theorem 2.1 and Lemma 2.2 from Section 2.1. First we calculate the second summand of (5.7), the probability that the branching process continues forever, and find that it is proportional to δ_n in the supercritical phase, and 0 in the subcritical phase.

Lemma 5.4. *If $\tau_n \downarrow 1$, then $\mathbb{P}[B_n = \infty] = 0$. If $\tau_n \uparrow 1$, then*

$$\mathbb{P}[B_n = \infty] \sim 4\delta_n. \tag{5.8}$$

Proof. If $\tau_n \downarrow 1$, the extinction probability is 1 by Theorem 2.3, so assume that $\tau_n \uparrow 1$. Let $\eta_n = \mathbb{P}[\beta_n = \infty]$. Theorem 2.3 implies that $p_n(1-\eta_n) = 1-\eta_n$, so by (5.5) and Lemma 2.2,

$$(1-\eta_n)\Lambda_n'(1) = \Lambda_n'(1-\eta_n) \overset{(2.2)}{=} \Lambda_n'(1) - \eta_n\Lambda_n''(1) + \frac{1}{2}\eta_n^2\Lambda_n'''(1) + O(\eta_n^3),$$

which gives us

$$\Lambda_n''(1) - \Lambda_n'(1) = \frac{1}{2}\eta_n\Lambda_n'''(1) + O(\eta_n^2). \tag{5.9}$$

The left hand side of this equation is $Q_n(1)$. Since $Q_n(\tau_n) = 0$ by assumption and $\delta_n \to 0$, we can use Lemma 2.2 to calculate this value. We first note that $Q_n(z)$ and $Q_n'(z)$ can be written as

$$\begin{aligned} Q_n(z) &= z^2\Lambda_n''(z) - z\Lambda_n'(z), \\ Q_n'(z) &= z^2\Lambda_n'''(z) + z\Lambda_n''(z) - \Lambda_n'(z) \\ &\overset{(5.2)}{=} z^2\Lambda_n'''(z) + \frac{Q_n(z)}{z}. \end{aligned} \tag{5.10}$$

Then, by Lemma 2.2 and (5.3),

$$\begin{aligned} Q_n(\tau_n) &= Q_n(1) - \delta_n Q_n'(1) + \delta_n^2 Q_n''(1) + O(\delta_n^3) \\ &\overset{(5.10)}{=} Q_n(1) - \delta_n(\Lambda'''(1) + Q_n(1)) + O(\delta_n^2), \end{aligned}$$

so

$$\begin{aligned} Q_n(1) &= \frac{1}{1-\delta_n}\left(\delta_n\Lambda'''(1) + O(\delta_n^2)\right) \\ &= \delta_n\Lambda'''(1) + O(\delta_n^2). \end{aligned} \tag{5.11}$$

Equations (5.9) and (5.11) imply that

$$\delta_n + O(\delta_n^2) = \frac{1}{2}\eta_n + O(\eta_n^2),$$

so

$$\mathbb{P}\left[\beta_n = \infty\right] \sim 2\delta_n. \tag{5.12}$$

Since $B_n = \infty$ if and only if $\beta_n^{(1)} = \infty$ or $\beta_n^{(2)} = \infty$, Equation (5.8) follows. □

We then have to find $\mathbb{P}\left[k \le B_n < \infty\right]$, and for this we will use Theorem 2.1. We recall that the functions $p_n(z)$ and $q_n(z)$ are related by (5.6). In view of Theorem 2.1, taking $\phi = p_n$ and $y = q_n$, we let τ_n be such that

$$p_n(\tau_n) - \tau_n p_n'(\tau_n) = 0.$$

This definition of τ_n is equivalent to (5.1), and can also be expressed as

$$\frac{\Lambda_n'(\tau_n)}{\tau_n\Lambda_n''(\tau_n)} = 1. \tag{5.13}$$

We define

$$\rho_n = \frac{\tau_n}{p_n(\tau_n)},$$

which is the radius of convergence of the generating function $q_n(z)$, and the location of the dominant singularity of $q_n(z)$, according to Theorem 2.1. The next lemma gives us a relation between ρ_n and δ_n.

Lemma 5.5. *Assume that $\tau_n \to 1$, and let ρ_n and δ_n be as before. Then*

$$\log \rho_n \sim c_2 \delta_n^2,$$

for a constant $c_2 = \frac{1}{2} + \frac{1}{2} \lim_{n\to\infty} \frac{\Lambda_n'''(1)}{\Lambda_n'(1)} > 0$.

Proof. Applying Lemma 2.2 to (5.3), we see that for all functions $f_n(z)$ satisfying the requirements of Lemma 2.2,

$$f_n(\tau_n) = f_n(1) - \delta_n f_n'(1) + \frac{1}{2}\delta_n^2 f_n''(1) + O(\delta_n^3). \tag{5.14}$$

Furthermore, using the equation

$$\log(1+x) = x - \frac{x^2}{2} + O(x^3), \tag{5.15}$$

we get

$$\begin{aligned} \log f_n(\tau_n) &\overset{(5.14)}{=} \log\left(f_n(1) - \delta_n f_n'(1) + \frac{\delta_n^2}{2} f_n''(1) + O(\delta_n^3)\right) \\ &= \log f_n(1) + \log\left(1 - \frac{\delta_n f_n'(1)}{f_n(1)} + \frac{\delta_n^2 f_n''(1)}{2 f_n(1)} + O(\delta_n^3)\right) \\ &\overset{(5.15)}{=} \log f_n(1) - \delta_n \frac{f_n'(1)}{f_n(1)} + \frac{\delta_n^2}{2}\left(\frac{f_n''(1)}{f_n(1)} - \frac{f_n'(1)^2}{f_n(1)^2}\right) + O(\delta_n^3). \end{aligned} \tag{5.16}$$

The function $\Lambda_n'(z)$ satisfies the conditions of Lemma 2.2. Hence,

$$\begin{aligned} \log \rho_n &\overset{(5.13)}{=} \log \frac{\tau_n \Lambda_n'(1)}{\Lambda_n'(\tau_n)} \\ &= \log \tau_n + \log \Lambda_n'(1) - \log \Lambda_n'(\tau_n) \\ &\overset{(5.16)}{=} \delta_n \left(\frac{\Lambda_n''(1)}{\Lambda_n'(1)} - 1\right) + \frac{\delta_n^2}{2}\left(\frac{\Lambda_n''(1)^2}{\Lambda_n'(1)^2} - \frac{\Lambda_n'''(1)}{2\Lambda_n'(1)} - 1\right) + O(\delta_n^3). \end{aligned} \tag{5.17}$$

Because $\tau_n \to 1$ we might expect from (5.13) that $\Lambda_n''(1)/\Lambda_n'(1)$ is close to 1. Indeed, applying Lemma 2.2 to the function $x\Lambda_n''(x)$, we get

$$\Lambda_n''(1) = \tau_n \Lambda_n''(\tau_n) + \delta_n(\Lambda_n''(1) + \Lambda_n'''(1)) + O\left(\delta_n^2\right),$$

and using instead the function $\Lambda_n'(z)$, we get

$$\Lambda_n'(1) = \Lambda_n'(\tau_n) + \delta_n \Lambda_n''(1) + O\left(\delta_n^2\right).$$

Thus

$$\begin{aligned}\frac{\Lambda_n''(1)}{\Lambda_n'(1)} &= \frac{\tau_n\Lambda_n''(\tau_n)+\delta_n(\Lambda_n''(1)+\Lambda_n'''(1))+O(\delta_n^2)}{\Lambda_n'(\tau_n)+\delta_n\Lambda_n''(1)+O(\delta_n^2)}\\ &= \frac{1}{\Lambda_n'(\tau_n)}\frac{\tau_n\Lambda_n''(\tau_n)+\delta_n(\Lambda_n''(1)+\Lambda_n'''(1))+O(\delta_n^2)}{1+\delta_n\frac{\Lambda_n''(1)}{\Lambda_n'(\tau_n)}+O(\delta_n^2)}\\ &= \frac{1}{\Lambda_n'(\tau_n)}\Bigg(\tau_n\Lambda_n''(\tau_n)+\\ &\qquad \delta_n\left(\Lambda_n''(1)+\Lambda_n'''(1)-\tau_n\Lambda_n''(\tau_n)\frac{\Lambda_n''(1)}{\Lambda_n'(\tau_n)}\right)+O(\delta_n^2)\Bigg).\end{aligned}$$

Using (5.13), and the fact that $\Lambda_n(\tau_n)=\Lambda_n(1)+O(\delta_n)$,

$$\begin{aligned}\frac{\Lambda_n''(1)}{\Lambda_n'(1)} &= 1+\delta_n\left(\frac{\Lambda_n''(1)+\Lambda_n'''(1)}{\Lambda_n'(\tau_n)}-\frac{\Lambda_n''(1)}{\Lambda_n'(1)}\right)+O(\delta_n^2)\\ &= 1+\delta_n\frac{\Lambda_n'''(1)}{\Lambda_n'(1)}+O(\delta_n^2),\end{aligned}$$

so by (5.17),

$$\log\rho_n=\frac{\delta_n^2}{2}\left(1+\frac{\Lambda_n'''(1)}{\Lambda_n'(1)}\right)+O(\delta_n^3),$$

which completes the proof. □

Lemma 5.6. *The probability that one branch of the branching process dies out after producing at least k vertices is*

$$\mathbb{P}\left[k\le\beta_n<\infty\right]\sim c_3e^{-c_2k\delta_n^2(1+o(1))}k^{-1/2}, \tag{5.18}$$

where $c_3>0$ is a constant, and $c_2>0$ is as in Lemma 5.5.

Proof. Recall that $q_k(n)$ is the probability that the branching process dies out after precisely k vertices have been produced. According to Theorem 2.1,

$$q_k(n)\sim c\rho_n^{-k}k^{-3/2}(1+O(k^{-1}))$$

for a constant c. Hence

$$\begin{aligned}\mathbb{P}\left[k\le\beta_n<\infty\right] &= \sum_{i\ge k}q_i(n)\sim\sum_{i\ge k}c\rho_n^{-i}i^{-3/2}(1+O(i^{-1}))\\ &\sim c\int_k^\infty\rho_n^{-x}x^{-3/2}(1+O(x^{-1}))dx\\ &\sim c\left[-\rho_n^{-x}x^{-1/2}\left(2+\frac{4}{3}\log\rho_n+o(x^{-1})\right)\right]_k^\infty\\ &= c\rho_n^{-k}k^{-1/2}\left(2+\frac{4}{3}\log\rho_n+o(k^{-1})\right).\end{aligned}$$

Now (5.18) follows from Lemma 5.5. □

Lemma 5.6 tells us the probability that one branch of the branching process dies out after k vertices have been created. The complete branching process has two branches, which produce $\beta_n^{(1)}$ and $\beta_n^{(2)}$ vertices respectively. We have

$$[k \le \beta_n^{(1)} < \infty] \wedge [\beta_n^{(2)} < \infty] \Rightarrow [k \le B_n < \infty]$$

and

$$[k \le B_n < \infty] \Rightarrow [k/2 \le \beta_n^{(1)} < \infty] \vee [k/2 \le \beta_n^{(2)} < \infty].$$

Hence we get the lower bound

$$\begin{aligned} \mathbb{P}[k \le B_n < \infty] &\overset{(5.12)}{\ge} \mathbb{P}\left[k \le \beta_n^{(1)} < \infty\right](1 - 2\delta_n) \\ &\sim c_3 e^{-c_2 k\delta_n^2(1+o(1))} k^{-1/2}, \end{aligned} \tag{5.19}$$

and the upper bound

$$\begin{aligned} \mathbb{P}[k \le B_n < \infty] &\le 2\mathbb{P}[k/2 \le \beta_n < \infty] \\ &\le 3c_3 e^{-\frac{1}{2}c_2 k\delta_n^2(1+o(1))} k^{-1/2}. \end{aligned} \tag{5.20}$$

Let C_n be the random variable denoting the number of vertices in the component containing a random edge in G_n. We will show that C_n and B_n behave very similarly.

Lemma 5.7. *Let α_n and γ_n be such that $k = \gamma_n n^{2/3}$ and $\delta_n = \alpha_n n^{-1/3}$. Suppose that $\gamma_n \ll |\alpha_n|$. Then there are constants c_4, c_5, c_6 and c_7 such that for large enough n,*

$$\mathbb{P}[C_n \ge k] \le c_4 e^{-c_5 k\delta_n^2(1+o(1))} k^{-1/2} + I_{\delta_n > 0} 4\delta_n(1 + o(1)). \tag{5.21}$$

and

$$\mathbb{P}[C_n \ge k] \ge c_6 e^{-c_7 k\delta_n^2(1+o(1))} k^{-1/2} + I_{\delta_n > 0} 4\delta_n(1 + o(1)). \tag{5.22}$$

Let $k_- = \gamma_n^- n^{2/3}$ and $k_+ = \gamma_n^+ n^{2/3}$, where $\gamma_n^- \le \gamma_n^+ \ll \alpha_n$. Then there is a positive constant c_8 such that for large enough n,

$$\mathbb{P}[k_- < C_n < k_+] \le c_8 e^{-c_2 \alpha^2 \gamma_-(1+o(1))} k_-^{-1/2}. \tag{5.23}$$

Proof. There are two problems which can cause B_n and C_n to differ. The first is the fact that in the random graph cycles can be formed, whereas this does not happen in the ordinary branching process. The second problem is that in the branching process the probability that a vertex has, say, i children

remains the same throughout the process. When exposing the component in the random graph, this is not true, since it depends on how many vertices of degree $i+1$ we have exposed so far. We will show that both of these factors have a negligible effect as long as the number of vertices exposed is not too large.

We first consider the possibility of cycles forming. Suppose that k vertices have already been exposed in $\mathcal{C}$. We first choose a half-edge w_1 in a partially exposed vertex, and then a half-edge w_2 uniformly at random from all unexposed half-edges. The probability that w_2 is in a partially exposed vertex is then $O\left(\frac{k}{n}\right)$.

On the other hand, let $X_{m,i}$ denote the number of vertices of degree i among the first m vertices exposed. These m vertices are picked at random from the total of n vertices, with every vertex being chosen with probability proportional to its degree. Since we consider the case that m is asymptotically small compared to n, the distribution of $X_{m,i}$ approaches a binomial distribution $\mathrm{Bin}\left(m, \frac{i\lambda_i}{d}\right)$ when n tends to infinity. Let b_m be the degree of the mth vertex. We obtain the upper bound

$$\begin{aligned}\mathbb{P}\left[b_m = i\right] &= \frac{ni\lambda_i - iX_{m,i}}{dn - \sum_j jX_{mj}} \\ &\leq \frac{i\lambda_i}{d}\frac{1}{1-\frac{1}{dn}\sum_j jX_{m,j}} \\ &= \frac{i\lambda_i}{d}\left(1+O\left(\frac{k}{n}\right)\right),\end{aligned}$$

when $m \leq k$. The probability that the mth vertex exposed has degree i is therefore $(1+o(1))\frac{i\lambda_i}{d}$ when m is small. Hence,

$$\mathbb{E}\left[X_{m,i}\right] = (1+o(1))\frac{mi\lambda_i}{d},$$

while the standard deviation is roughly $\sigma_{m,i} = \sqrt{m\frac{i\lambda_i}{d}\left(1-\frac{i\lambda_i}{d}\right)}$. Using Chernoff bounds, we can show that there is a constant c, such that with probability $1-o(n^{-2})$ we have

$$|X_{m,i} - \mathbb{E}\left[X_{m,i}\right]| \leq c\sigma_{m,i}\sqrt{\log n}. \tag{5.24}$$

Hence, with probability $1-o(n^{-1})$ (5.24) holds for all $i = 1, \ldots, \psi(n)$, where $\psi(n)$ is the maximum degree. We can therefore condition on this holding.

Let K_n be the set of integers i with $1 \leq i \leq \psi(n)$ such that $m\frac{i\lambda_i}{d} \geq \log^2 n$. By (5.24), for every $i \in K_n$, $X_{m,i} = (1+o(1))\mathbb{E}\left[X_{m,i}\right] = (1+o(1))\frac{mi\lambda_i}{d}$, and

for every $i \notin K_n$, $X_{m,i} \leq c' \log n$ for some constant c'. We therefore have

$$\begin{aligned}\sum_i iX_{m,i} &= \sum_{i \in K_n} (1+o(1)) \frac{mi^2\lambda_i}{d} + O\left(\sum_{i \notin K_n} \log n\right) \\ &= (1+o(1)) \frac{m}{d} \sum_i i^2\lambda_i + O\left(\psi(n) \log n\right) \\ &= (1+o(1)) \frac{m}{d} (\Lambda''(1) + \Lambda'(1)).\end{aligned}$$

Then we can also find a lower bound on the probability that the mth vertex has degree i.

$$\begin{aligned}\mathbb{P}\left[b_m = i | i \in K_n\right] &= \frac{ni\lambda_i - iX_{m,i}}{dn - \sum_j jX_{m,j}} \\ &= \frac{i\lambda_i}{d} \frac{1 - \frac{(1+o(1))mi}{nd}}{1 - \frac{m}{d^2n}(\Lambda''(1)+d)(1+o(1))} \\ &= \frac{i\lambda_i}{d}\left(1 + O\left(\frac{k}{n}\right)\right), \qquad (5.25)\end{aligned}$$

for $1 \leq m \leq k$, whereas

$$\begin{aligned}\mathbb{P}\left[b_m = i | i \notin K_n\right] &= \frac{ni\lambda_i - iX_{m,i}}{dn - \sum_j jX_{m,j}} \\ &\geq \frac{ni\lambda_i - \frac{mi\lambda_i}{d} - c\sqrt{\frac{mi\lambda_i}{d}\log n}}{dn} \\ &\geq \frac{ni\lambda_i - c' \log^2 n}{dn} \\ &= \frac{i\lambda_i}{d} - \frac{c' \log^2 n}{dn}.\end{aligned}$$

We will denote the branching process defined in Section 5.2 by $\mathcal{B}$. Thus we may say that if we use the branching process $\mathcal{B}$ to approximate the component exposure process $\mathcal{C}$, then at every step there is a chance that we choose the "wrong" degree. We therefore introduce a modified branching process $\overline{\mathcal{B}}$ as follows. In $\mathcal{B}$ an edge gets i children with probability $p_i = \frac{(i+1)\lambda_{i+1}}{d}$. If $i \in K_n$, then an edge in $\overline{\mathcal{B}}$ gets i children with probability $\overline{p}_i = p_i(1+\varepsilon_i)$, where $\varepsilon_i = \varepsilon_i(n)$ are error terms depending on n, such that $|\varepsilon_i| \ll |\delta|$. If $i \notin K_n$, then an edge in $\overline{\mathcal{B}}$ gets i children with probability $\overline{p}_i = p_i + \varepsilon_i$, where again $\varepsilon_i = \varepsilon_i(n)$ depends on n, and $|\varepsilon_i| \leq \frac{c' \log^2 n}{n}$. Note that if $i \notin K_n$, then the error term is not relative to p_i. By choosing the functions ε_i appropriately, we can make sure that the process $\overline{\mathcal{B}}$ becomes either an upper bound or a

lower bound of the exposure process $\mathcal{C}$, and we will show that as long as the stated bounds are satisfied, the behaviour of $\overline{\mathcal{B}}$ does not differ much from the behaviour of $\mathcal{B}$. We then obtain the probability generating function for the number of children of an edge in the modified branching process as

$$\begin{aligned}\overline{p}_n(z) &= \sum_i \overline{p}_i z^i = \sum_{i\in K_n} p_i(1+\varepsilon_i)z^i + \sum_{i\notin K_n}(p_i+\varepsilon_i)z^i \\ &= p_n(z) + \sum_{i\in K_n} p_i\varepsilon_i z^i - \sum_{i\notin K_n}\varepsilon_i z^i.\end{aligned}$$

We are interested in the behaviour of $\overline{p}_n(z)$ when $z \in (1-2\delta, 1+2\delta)$. In this interval $z^{\psi(n)} = 1 + o(1)$ since $\psi(n) = o(n^{1/4})$. Hence

$$\left|\sum_{i\in K_n} p_i\varepsilon_i z^i\right| = o(\delta)\left|\sum_{i\in K_n} p_i z^i\right| = o(\delta),$$

and

$$\left|\sum_{i\notin K_n}\varepsilon_i z^i\right| \le (1+o(1))\psi(n)\frac{c'\log n}{dn} = o(n^{-3/4}).$$

We can therefore write $\overline{p}_n(z) = p_n(z) + c_n(z)$, where $c_n(z)$ is a function such that $|c_n(z)| = o(\delta)$ when $z \in (1-2\delta, 1+2\delta)$. Likewise we can show that $|\sum_{i\in K_n} i p_i\varepsilon_i z^{i-1}| = o(\delta_n)$ and $\left|\sum_{i\notin K_n} i\varepsilon_i z^i\right| = o(n^{-1/2})$, so that also $|c_n'(z)| = o(\delta)$. We let $\overline{\tau}_n$ be defined such that $\overline{p}_n(\overline{\tau}_n) - \overline{\tau}_n\overline{p}_n'(\overline{\tau}_n) = 0$, and let $\overline{\delta}_n = 1 - \overline{\tau}_n$. Let $\zeta_n = \overline{\tau}_n - \tau_n$. Then

$$\begin{aligned}0 &= \overline{p}_n(\overline{\tau}_n) - \overline{\tau}_n\overline{p}_n'(\overline{\tau}_n) \\ &= (p_n(\overline{\tau}_n) + c_n(\overline{\tau}_n)) - \overline{\tau}_n\left(p_n'(\overline{\tau}_n) + c_n'(\overline{\tau}_n)\right) \\ &= p_n(\tau_n+\zeta_n) - (\tau_n+\zeta_n)p_n'(\tau_n+\zeta_n) + o(\delta_n) \\ &= p_n(\tau_n) + \zeta_n p_n'(\tau_n) - (\tau_n+\zeta_n)(p_n'(\tau_n) + \zeta_n p_n''(\tau_n)) + O(\zeta_n^2) + o(\delta_n) \\ &= -\zeta_n p_n''(1) + O(\zeta_n^2) + o(\delta_n),\end{aligned}$$

by Lemma 2.2. Hence $\zeta_n = o(\delta_n)$, so $\overline{\delta}_n \sim \delta_n$. Let $\overline{B}_n$ be the random variable denoting the number of vertices generated in $\overline{\mathcal{B}}$, and let $\overline{q}_n(z)$ be the corresponding probability generating function. Then $\overline{q}_n(z)$ is given implicitly by $\overline{q}_n(z) = z\overline{p}_n(\overline{q}_n(z))$. According to Theorem 2.1, the dominant singularity of $\overline{q}_n(z)$ is $\overline{\rho}_n = \frac{\overline{\tau}_n}{\overline{p}_n(\overline{\tau}_n)}$. Lemma 5.5 states that $\log\rho_n \sim c_2\delta_n^2$; we can similarly calculate that $\log\overline{\rho}_n \sim c_2\delta^2$. Hence, by Theorem 2.1,

$$\overline{q}_k \sim c\overline{\rho}_n^{-k}k^{-3/2} = ce^{-c_2k\log\rho_n}k^{-3/2} = ce^{-c_2k\delta_n^2(1+o(1))}k^{-3/2}, \tag{5.26}$$

where c is a positive constant. Furthermore, as in Lemma 5.4, the probability that $\overline{\mathcal{B}}$ does not die out is 0 if $\delta_n < 0$, and if $\delta_n > 0$, then $\mathbb{P}\left[\overline{B}_n = \infty\right] \sim 4\overline{\delta}_n \sim 4\delta_n$.

Instead of approximating $\mathcal{C}$ by the branching process $\mathcal{B}$, we will approximate it by $\overline{\mathcal{B}}$, choosing the functions $\varepsilon_i(n)$ appropriately. In $\mathcal{C}$ the probability that the next vertex chosen has degree i depends on the previously exposed vertices, while in $\overline{\mathcal{B}}$, the probabilities are the same all the time. In (5.25) we gave bounds for the deviations of the probabilities between $\mathcal{B}$ and $\mathcal{C}$. In the definition of $\overline{\mathcal{B}}$, we assumed that the deviations ε_i are asymptotically smaller than δ, while in (5.25) the deviations were found to be $O\left(\frac{k}{n}\right)$. By assumption, $\frac{k}{n} = \gamma_n n^{-1/3} \ll \alpha_n n^{-1/3} = \delta_n$, so in $\overline{\mathcal{B}}$, the error terms ε_i can be chosen in such a way that $\overline{\mathcal{B}}$ becomes either a lower bound of $\mathcal{C}$, or an upper bound. We can therefore derive (5.21) in the same manner as (5.20), and (5.22) in the same manner as (5.19). As for (5.23), we use (5.26) to obtain

$$\begin{aligned}\mathbb{P}\left[k_- < C_n < k_+\right] &= \sum_{k=k_-}^{k_+} \mathbb{P}\left[C_n = k\right] \sim \int_{k_-}^{k_+} ce^{-c_2 x\delta^2(1+o(1))} x^{-3/2} dx \\ &\leq 2ce^{-c_2 k_- \delta^2(1+o(1))} {k_-}^{-1/2} \\ &= 2ce^{-c_2\alpha^2\gamma_-(1+o(1))} k_-^{-1/2}.\end{aligned}$$

□

5.4 The phase transition

Using the lemmas of the previous section, we can now prove Theorem 5.2.

5.4.1 The subcritical case

We first consider case (i) of Theorem 5.2, that $\alpha_n = \delta_n n^{1/3} \to -\infty$. We want to show that the largest component in G_n a.a.s. has $o(n^{2/3})$ vertices.

Let $k = cn^{2/3}$ for some constant c. Since clearly $c \ll |\alpha_n|$, (5.21) implies that

$$\mathbb{P}\left[C_n \geq k\right] \leq \frac{c_4}{\sqrt{c}} e^{-cc_5\alpha_n^2} n^{-1/3}, \tag{5.27}$$

where $c_4, c_5 > 0$. Let X_k be the number of vertices in components of order greater than k, and let A_k be the event that there is a component of order at least k. Then, by (5.27) and Markov's inequality,

$$\mathbb{P}\left[A_k\right] = \mathbb{P}\left[X_k \geq k\right] \leq \frac{\mathbb{E}\left[X_k\right]}{k} \sim \frac{n\mathbb{P}\left[C_n \geq k\right]}{k} \leq \frac{c_4}{c^{3/2}} e^{-cc_5\alpha_n^2} \to 0.$$

There is therefore a.a.s. no component in G_n^* with more than $cn^{2/3}$ vertices, for every positive constant c.

5.4.2 The supercritical case

Now we consider the supercritical phase, when $\alpha_n = \delta_n n^{1/3} \geq c_1 \log n$. We call a component *large* if it has $\gg n^{2/3}$ vertices and small if it has $o(n^{2/3})$ vertices. We will prove firstly that a.a.s. every component is either large or small, secondly that there is a.a.s. at least one large component, and thirdly that there is a.a.s. only one large component.

Lemma 5.8. *Let $\omega(n)$ be a function which tends to infinity as $n \to \infty$, but such that $\omega(n) = o(\log n)$. There is a constant c_9, such that if $\alpha_n \geq c_9 \log n$, then the probability that G_n^* contains a component of order between $k_- = n^{2/3}/\log n$ and $k_+ = n^{2/3}\omega(n)$ is $O(n^{-1})$.*

Proof. If v is a vertex, we let $C(v)$ be the component containing v. Then, according to (5.23),

$$\begin{aligned} \mathbb{P}\left[k_- \leq |C(v)| \leq k_+\right] &\leq c_8 e^{-c_2\alpha_n^2/\log n} k_-^{-1/2} \\ &\leq c_8 n^{-c_2 c_9^2} n^{-1/3}\sqrt{\log n}. \end{aligned}$$

Let X be the number of vertices contained in components of order between k_- and k_+, and let A be the event that there is at least one such component. Then, by Markov's inequality,

$$\begin{aligned} \mathbb{P}\left[A\right] = \mathbb{P}\left[X \geq k_-\right] \leq \frac{\mathbb{E}\left[X\right]}{k_-} &= \frac{n\mathbb{P}\left[k_- \leq |C(v)| \leq k_+\right]}{k_-} \\ &\leq c_8 n^{-c_2 c_9^2} \log^{3/2} n. \end{aligned}$$

Clearly c_9 can be chosen so large that $\mathbb{P}\left[A\right] = O(n^{-1})$. □

We assume that c_1 in Theorem 5.2 satisfies $c_1 \geq c_9$. Thus, by Lemma 5.8 we know that a.a.s. every component in G_n^* is either large or small. We will now show that there is a.a.s. at least one large component in G_n^*. We will do this by considering a subgraph H of G_n^*, which can also be viewed as a random graph with degree sequence $\mathcal{D}_0'$, where $\mathcal{D}_0'$ is some asymptotic degree sequence.

Let us consider the configuration model $\mathcal{C}_n$. We know that the perfect matching $\mathcal{P}_n$ can be constructed greedily. We will construct it in the following way: for some m, we first choose m pairs of half-edges randomly, and label them $e_1, \ldots, e_m$. Let $L_0' = L_n \setminus \{e_1, \ldots, e_m\}$ be the set of unmatched half-edges, and let $\mathcal{D}_0'$ be the degree sequence of L_0'. Then we choose a random

perfect matching, $\mathcal{P}_0'$, of L_0', and let $\mathcal{C}_0'$ be the random configuration consisting of the set L_0' of half-edges and the matching $\mathcal{P}_0'$. Let H be the underlying multigraph of $\mathcal{C}_0'$. Then H is a subgraph of G_n^*.

Each of the pairs of half-edges e_i, with $1 \leq i \leq m$, is chosen in the following way. We first choose a half-edge uniformly at random from the set of yet unmatched half-edges in vertices of degree at least 3. The second half-edge is chosen uniformly at random from all unmatched half-edges. This is a valid way to generate the random matching $\mathcal{P}_n$, since, as we stated in Section 1.5.1, the first half-edge in every pair can be chosen in an arbitrary manner, as long as the second half-edge is chosen uniformly at random. Moreover, we will make sure that we always have $m = o(n)$, and since, by assumption, $\lambda_j^* > 0$ for some $j > 2$, the set of vertices of degree at least 3 will not be exhausted.

Recall that $\alpha_n = \delta_n n^{1/3}$. We write $\alpha_n = \alpha(\mathcal{D}_n)$, such that α is a function of the degree sequence $\mathcal{D}_n$. In the process explained in the previous paragraph we do not fix m beforehand, but we choose the pairs $e_1, e_2, \ldots$ one by one and remove them from L_n. Whenever an edge is removed from L_n in this manner, the value of $Q_n(z)$ decreases. Indeed, by assumption, at least one of the end-vertices of e_i, for any i, has degree 3 or greater. In the case that the end-vertices have degree 1 and 3 respectively, the value of $Q_n(1)$ decreases by $\frac{2}{n}$. Any other combination of degrees causes $Q_n(1)$ to decrease by a greater amount.

We know from (5.4) that $Q_n(1)$ and δ are proportional. Hence δ and α similarly decrease whenever a pair of half-edges is removed from L_n. It is clear that removing sufficiently many pairs of half-edges will cause $Q_n(1)$, and thereby α, to become negative. We will continue to remove pairs of half-edges until the degree sequence $\mathcal{D}'$ of the remaining set is such that $\alpha(\mathcal{D}') \leq \log^{3/8} n$. Since the maximum degree is less than $n^{1/4}$ by assumption, the value of α decreases by at most $O(n^{-3/4})$ for every pair removed; hence we will have $\alpha_0' := \alpha(\mathcal{D}') \sim \log^{3/8} n$. The graph H is then a random multigraph with degree sequence $\mathcal{D}_0'$.

Set $\gamma_- = \log^{-9/10} n$ and $\gamma_+ = \log^{2/5} n$, and let $k_- = n^{2/3}\gamma_-$ and $k_+ = n^{2/3}\gamma_+$. We will show that H a.a.s. contains a component of order at least k_-. This implies that G_n^* a.a.s. contains a component of order at least k_-, and Lemma 5.8 then implies that G_n^* a.a.s. has a component of order asymptotically greater than $n^{2/3}$.

We let C_0' be the random variable denoting the order of the component containing a specified vertex of H. We let A_k' be the event that there is at least one component in H of order at least k, and we let X_k' be the number of vertices which are contained in such components. Then, using (5.21), we

get

$$\mathbb{P}\left[C_0' \geq k_+\right] \leq c_4 e^{-c_5\gamma_+(\alpha_0')^2(1+o(1))}k_+^{-1/2} + 4\alpha_0' n^{-1/3}(1+o(1)) \sim 4\alpha_0' n^{-1/3}.$$

Thus

$$\mathbb{P}[A_{k_+}] = \mathbb{P}[X_{k_+}' \geq k_+] \leq \frac{n}{k_+}\mathbb{P}\left[C_0' \geq k_+\right] \leq \frac{4\alpha_0'}{\gamma_+} = 4\log^{-1/40} n = o(1).$$

We can therefore condition on $\overline{A_{k_+}}$, namely that H does not contain any components of order k_+ or greater.

By (5.22),

$$\begin{aligned}
\mathbb{E}[X_{k_-}'] &= n\mathbb{P}\left[C_0' \geq k_-\right] \\
&\geq n\left(c_6 e^{-c_7\gamma_-(\alpha_0')^2(1+o(1))}k^{-1/2} + 4\alpha_0' n^{-1/3}(1+o(1))\right) \\
&\sim n^{2/3}\left(c_6\gamma_-^{-1/2}e^{-c_7(1+o(1))\log^{-0.15} n} + 4\log^{3/8} n\right) \\
&\sim n^{2/3}\left(c_6\log^{9/20} n + 4\log^{3/8} n\right) \sim c_6 n^{2/3}\log^{9/20} n.
\end{aligned}$$

Suppose that v is a vertex in a component of order at least k_-. Since we condition on $\overline{A_{k_+}}$, the component containing v has at most k_+ vertices. The expected number of vertices $w \neq v$, such that w is also contained in a component of order at least k_- is therefore bounded by $k_+ + \mathbb{E}[X_{k_-}']$. Since $k_+ \ll \mathbb{E}[X_{k_-}']$,

$$\begin{aligned}
\mathbb{E}[X_{k_-}'(X_{k_-}' - 1)] &\leq \mathbb{E}[X_{k_-}']\left(k_+ + \mathbb{E}[X_{k_-}']\right) \\
&= \mathbb{E}[X_{k_-}']^2(1+o(1)),
\end{aligned}$$

so by Chebyshev's inequality, a.a.s. $X_{k_-}' \sim \mathbb{E}[X_{k_-}']$. We conclude that H, and thereby G_n^*, a.a.s. contains a component of order at least k_-. Hence, according to Lemma 5.8, G_n^* a.a.s. contains at least one large component.

It remains to prove that there is just one such component in G_n^*. We will do this by deleting a certain number of edges in the same way as in the previous step, but now we fix $m = \lceil n^{2/3}\log n\rceil$. Furthermore, this time we will choose the edges $e_1, \ldots, e_m$ uniformly at random. We let as before $L_0' = L \setminus \{e_1, \ldots, e_m\}$, and $\mathcal{P}_0'$ be a random perfect matching of L_0'. For $1 \leq i \leq m$, we let $L_i' = L_{i-1}' \cup \{e_i\}$, and let $\mathcal{D}_i'$ be the degree sequence of L_i'. Furthermore we let $\mathcal{P}_i'$ be the perfect matching of L_i' consisting of the pairs in $\mathcal{P}_0'$ together with the pairs $e_1, \ldots, e_i$. We let H_i be the underlying multigraph of the configuration $\mathcal{C}_i'$, which consists of L_i' and $\mathcal{P}_i'$. Then H_i can be considered a random graph with given degree sequence $\mathcal{D}_i'$, and $G_n^* = H_m$.

By assumption $\alpha_n = \alpha(\mathcal{D}'_m) \geq c_1 \log n$. We will first show that $\alpha'_0 = \alpha(\mathcal{D}'_0) \asymp \alpha_n$. Let λ'_i be the proportion of vertices in H_0 having degree i. Let $Q_n(x) = \sum_i i(i-2)\lambda_i x^i$ and $Q_0(x) = \sum_i i(i-2)\lambda'_i x^i$. Recall that $\tau_n = 1 - \delta_n$, and that $Q_n(\tau_n) = 0$. We let τ'_0 be a number such that $Q_0(\tau'_0) = 0$, and we let $\delta'_0 = 1 - \tau'_0$. We have to show that $\delta'_0 \asymp \delta_n$.

We let $D(x) = Q(x) - Q_0(x) = \sum_i i(i-2)(\lambda_i - \lambda'_i)x^i$. Suppose first that the difference between the degree sequence of $H_m = G^*_n$, and the degree sequence of H_0 is that precisely one vertex has degree d in H_0, but degree $d+1$ in H_m, while all other vertices have unchanged degree. (Ignore for a moment that this is impossible.) Then

$$\begin{aligned}\sum_i i(i-2)(\lambda_i - \lambda'_i)\tau_n^i &= \frac{1}{n}(-d(d-2)\tau_n^d + (d+1)(d-1)\tau_n^{d+1}) \\ &= \frac{\tau_n^d}{n}(2d - 1 + (1-d^2)\delta_n).\end{aligned}$$

Now we remember that the difference between H_m and H_0 is more substantial, namely that m edges have been added to H_0 to obtain H_m. Let m_i be the number of times any vertex gets its degree increased from i to $i+1$ in this process. Then

$$\begin{aligned}D(\tau_n) &= \sum_i i(i-2)(\lambda_i - \lambda'_i)\tau_n^i \\ &= \frac{1}{n}\sum_i (2i - 1 + (1-i^2)\delta_n)m_i\tau_n^i,\end{aligned}$$

and $2m = \sum_i m_i$. Since $m = n^{2/3}\log n$, $\tau_n \to 1$ and $\delta_n \to 0$, we see that $D(\tau_n) \sim c_{10}\frac{m}{n} = c_{10}n^{-1/3}\log n$ for a constant $c_{10} > 0$. Hence

$$Q_0(\tau_n) = Q(\tau_n) - D(\tau_n) = -c_{10}n^{-1/3}\log n. \tag{5.28}$$

Let $\Delta\tau = \tau_n - \tau'_0$. By Lemma 2.2,

$$Q_0(\tau_n) = Q_0(\tau'_0) + \Delta\tau Q'_0(\tau'_0) + \Delta\tau^2 Q''_0(\tau'_0) + O(\Delta\tau^3). \tag{5.29}$$

Equations (5.28) and (5.29) imply that

$$\Delta\tau Q'_0(\tau'_0) + O(\Delta\tau^2) = -c_{10}n^{-1/3}\log n,$$

so $\Delta\tau \asymp -n^{-1/3}\log n$. Since the difference between τ_n and τ'_0 is in the order of $n^{-1/3}\log n$, we get that $\alpha'_0 \sim (c_1 - c_{10})\alpha_n$. We assume that c_1 is so large that $c_1 - c_{10} \geq c_9$, where c_9 is the constant in Lemma 5.8.

Let $\alpha'_i = \alpha(\mathcal{D}'_i)$. Since $\alpha'_{i+1} \geq \alpha'_i$, we have $\alpha'_i \geq c_9 \log n$ for all $i = 0, \ldots, m$. Thus, by Theorem 5.8, the probability that there is a component

with between $n^{2/3}/\log n$ and $n^{2/3}$ in any of the graphs $H_0, \dots, H_m$ is bounded by $n^{2/3-1}\log n = o(1)$. It follows that every large component in H_i for $i = 1, \dots, m$ must contain some large component in H_{i-1}, and hence every large component in H_m must contain at least one of the large components in H_0.

Let $C_1, \dots, C_l$ be the large components in H_0. We must prove that these components a.a.s. are contained in one component in G_n^*. Recall that m edges were removed from G_n^* to obtain H_0. We will show that for every pair (C_i, C_j) of large components in H_0, it is very likely that one of the edges removed from G_n^* joins two vertices in C_i and C_j to each other.

Let $E = \{e_1, \dots, e_m\}$, and let M be the set of vertices in G_n^* which are incident to one of the edges in E. Suppose that v is a vertex of degree i in G_n^*. The probability that e_j is incident to v is $\frac{i}{dn}$. It follows that the expected number of vertices of degree i which are incident to one of the edges in E is $m\frac{i\lambda_i}{d}$. The expected number of edges in H_0 which are incident to a vertex in M is then $\sum_i (i-1)m\frac{i\lambda_i}{d} = \frac{m}{d}\sum_i i(i-1)\lambda_i$. Because $\mathcal{D}$, and therefore also $\mathcal{D}_0$, is well-behaved, the sum tends to a constant, so the expected number of edges in H_0 incident to a vertex in M is $c_{11}m$ for some constant $c_{11} > 0$.

Now let e be any edge in H_0, and suppose we expose the component containing e as explained in Section 5.2. At any point in the exposure process the probability that the next vertex is a vertex in M is $\frac{c_{11}m}{dn} \asymp n^{-1/3}\log n$. Let $\omega(n) \to \infty$ be such that each of the components $C_1, \dots, C_l$ contains at least $n^{2/3}\omega(n)$ vertices. The expected number of vertices in M among the first $n^{2/3}\omega(n)$ vertices exposed in any component is $\Theta(n^{1/3}\omega(n)\log n)$. Furthermore, the distribution of the number of such vertices tends to a binomial distribution $\mathrm{Bin}(n^{2/3}\omega(n), \frac{c_{11}}{d}n^{-1/3}\log n)$ as $n \to \infty$, so we assert that for any $i = 1, \dots, l$, with probability $1 - o(n^{-1})$, the number of vertices in $M \cap C_i$ is at least $c_{12}n^{1/3}\omega(n)\log n$ for some constant $c_{12} > 0$.

There can be at most $n^{1/3}$ large components in H_0, so a.a.s. each of these components has at least $c_{12}n^{1/3}\omega(n)\log n$ vertices from M. Consider two components C_i and C_j in H_0 with $1 \le i, j \le l$, and let v be a vertex in $M \cap C_i$. The probability that one of the edges in E has v as one endpoint and its other endpoint in C_j is $\frac{|M\cap C_j|}{|M|} \ge \frac{c_{12}n^{1/3}\log n}{2n^{2/3}\log n} \asymp n^{-1/3}$. The probability that none of the vertices in $M \cap C_i$ is adjacent to a vertex in $M \cap C_j$ is therefore

$$\left(1 - \frac{1}{n^{1/3}}\right)^{c_{12}n^{1/3}\omega(n)\log n} \sim e^{-c_{12}\omega(n)\log n} = o(n^{-1}).$$

Since $l \le n^{1/3}$, the expected number of pairs of the components C_i, C_j with $1 \le i, j \le l$, which are *not* connected by one of the m edges, is then at most $n^{2/3}n^{-1} = o(1)$. Hence, a.a.s. all the components have joined to form a single

component. This concludes the proof of Theorem 5.2.

5.5 Comparision with $G(n,p)$

The parameter $\delta_n n^{1/3}$, or alternatively $Q_n(1)n^{1/3}$, appears to play the same role in Theorem 5.2 as λ does in Theorem 1.2. Thus Theorem 5.2 indicates that the phase transition in the random graph model $G_n(\mathcal{D})$ is similar to the phase transition in $G(n,p)$. We believe that part (iii) of Theorem 5.2 holds whenever $\delta_n n^{1/3} \to \infty$, which would make this theorem entirely analogous to Theorem 1.2.

Chapter 6

The critical behaviour of random digraphs

Let $D(n,p)$ be a random digraph with vertex set $[n] = \{1, 2, \ldots, n\}$, where each of the $n(n-1)$ possible arcs is included in the digraph, independently of each other, with probability $p = p(n)$. By a *component* of a digraph D we mean a maximal, strongly connected subgraph of D. A *complex* component is a component which is neither a single vertex nor a cycle.

The phase transition phenomenon in $D(n,p)$ was studied by Łuczak [38] and Karp [31]. (For an analogous result for a multiparameter generalisation of $D(n,p)$, see Łuczak and Cohen [39].) They proved that if $np \to c < 1$ and $\omega(n) \to \infty$, then a.a.s. $D(n,p)$ contains no complex components, and each component of $D(n,p)$ has fewer than $\omega(n)$ vertices. On the other hand, if $np \to c > 1$, then a.a.s. $D(n,p)$ contains a unique complex component on $\Theta(n)$ vertices, while all other components of $D(n,p)$ are cycles and isolated vertices. Karp [31] also considered the structure of $D(n,p)$ when $np = 1 + \varepsilon$ and $\varepsilon = \varepsilon(n)$ is a function which slowly tends to 0 as $n \to \infty$. The main result in this chapter states that the component structure of $D(n,p)$ changes when $np = 1 + \varepsilon$ and $\varepsilon = \varepsilon(n) = \Theta(n^{-1/3})$.

Theorem 6.1. *Let $np = 1 + \varepsilon$ such that $\varepsilon = \varepsilon(n) \to 0$, and let $\omega(n) \to \infty$.*

(i) If $n\varepsilon^3 \to -\infty$, then a.a.s. every component in $D(n,p)$ is either a vertex or a cycle of length at most $\omega(n)/|\varepsilon|$.

(ii) If $n\varepsilon^3 \to \infty$, then a.a.s. $D(n,p)$ contains a unique complex component, which has order $(4 + o(1))\varepsilon^2 n$, while every other component is either a vertex or a cycle of length at most $\omega(n)/\varepsilon$.

6.1 $D(n,p)$ and $G(n,p)$

In order to outline our argument, we need to introduce some notation. If v and w are two vertices, and V is a set of vertices in $D(n,p)$, then $d_V(v,w)$ denotes the length of the shortest path from v to w in $D(n,p)$ which visits only vertices in V; if no such path exists, we put $d_V(v,w)=\infty$. We also set

$$\overrightarrow{S}_d(V,v)=\{w : d_V(v,w)=d\}$$

and

$$\overleftarrow{S}_d(V,v)=\{w : d_V(w,v)=d\}.$$

Thus, $\overrightarrow{S}(V,v)=\bigcup_{d\geq 0}\overrightarrow{S}_d(V,v)$ and $\overleftarrow{S}(V,v)=\bigcup_{d\geq 0}\overleftarrow{S}_d(V,v)$ stand for the sets of all vertices that can be reached from v in V and all vertices in V which can reach v, respectively. In the case that V is the whole vertex set of D, we will suppress it and write $\overrightarrow{S}(v)=\overrightarrow{S}(D,v)$ and so on. The vertices in $\overrightarrow{S}(v)$ are called the *descendants* of v, and the vertices in $\overleftarrow{S}(v)$ are the *ancestors* of v. Note that v is an element of both $\overrightarrow{S}(v)$ and $\overleftarrow{S}(v)$, and that there may be other vertices which are both descendants and ancestors of v; these are precisely the vertices comprising the component containing v. By $S(v;n,p)$ we denote the random variable which counts the vertices in $\overrightarrow{S}(v)$ for a vertex v in $D(n,p)$; in a similar way $S_d(v;n,p)$ denotes the size of $\overrightarrow{S}_d(v)$ in $D(n,p)$. Clearly, the choice of v has no effect on the distribution of $S(v;n,p)$ and $S_d(v;n,p)$; we therefore let $S(n,p)$ and $S_d(n,p)$ be random variables with the same distribution as $S(v;n,p)$ and $S_d(v;n,p)$, respectively.

Note that for every integer k, the probability that $\overrightarrow{S}(v)$ has k elements in $D(n,p)$ equals the probability that v' is contained in a component of order k in $G(n,p)$. Indeed, let us assume that we want to find all vertices of $D(n,p)$ contained in $\overrightarrow{S}(v)$. A natural way of doing this is using a sequential search, say the breadth-first search. Whenever w is a vertex we have already determined is in $\overrightarrow{S}(v)$, and x is a vertex whose membership status in $\overrightarrow{S}(v)$ we do not know yet, we ask whether (w,x) is an arc in $D(n,p)$, and we get an affirmative answer with probability p, in which case we include x in $\overrightarrow{S}(v)$. Exposing the component containing v' in $G(n,p)$ can be done in the same way, and since the probability of an affirmative answer is the same in $G(n,p)$ and $D(n,p)$ whenever we ask about the presence of an edge, respectively, an arc, the probability of finding, say, k vertices by this process is the same in $D(n,p)$ and $G(n,p)$. This argument holds as long as, in $D(n,p)$, we do not check whether the arc (v,w) and the arc (w,v) are both contained in $D(n,p)$.

We shall often use the above observation to deduce properties of $D(n,p)$ from results on $G(n,p)$, whose structure is nowadays well studied and under-

stood. The following theorem from Łuczak [35] (see also Chapter 5 in [27]), which supplemented an earlier result of Bollobás [8] (see also [11]), can be considered an analogue of Theorem 6.1 for the undirected case. (This is a more precise version of Theorem 1.2.)

Theorem 6.2. *Let $np = 1 + \varepsilon$, such that $\varepsilon = \varepsilon(n) \to 0$ but $n\varepsilon^3 \to \infty$, and $k_0 = 2\varepsilon^{-2} \log n|\varepsilon|^3$.*

(i) If $n\varepsilon^3 \to -\infty$, then $G(n,p)$ a.a.s. contains no component of order greater than k_0. Moreover, a.a.s. each component of $G(n,p)$ is either a tree or contains precisely one cycle.

(ii) If $n\varepsilon^3 \to \infty$, then $G(n,p)$ a.a.s. contains exactly one component of order greater than k_0. This component a.a.s. has $(2+o(1))\varepsilon n$ vertices.

Let us observe that it follows from Theorem 6.2(i) that if v is a given vertex in $D(n,p)$, then the probability that $|\overrightarrow{S}(v)| \geq k_0$ is $o(1)$. However, this does not directly imply that $D(n,p)$ a.a.s. contains no vertices v such that $|\overrightarrow{S}(v)| \geq k_0$.

We shall also need some more technical results, which have been used in the studies of the critical behaviour of $G(n,p)$. We first remark that from the calculations used in the proof of Theorem 6.2(ii), contained mainly in [8] (see also [11], [27] and [34]), it follows that in the supercritical phase — that is when $np = 1 + \varepsilon$, where $n\varepsilon^3 \to \infty$ — the expected number of vertices in components of $G(n,p)$ larger than k_0 is $(2 + o(1))\varepsilon$, and, moreover, for every constant $\delta > 0$ the expected number of vertices which are contained in components of $G(n,p)$ whose sizes belong to $[0.1k_0, (2-\delta)\varepsilon n] \cup [(2+\delta)\varepsilon n, n]$ is $o(\varepsilon n)$. Consequently, the following holds.

Lemma 6.3. *If $np = 1 + \varepsilon$, where $\varepsilon = \varepsilon(n) \to 0$ but $n\varepsilon^3 \to \infty$, then the probability that a randomly chosen vertex v from $G(n,p)$ is contained in a component of order larger than $0.1k_0$ is $(2+o(1))\varepsilon$. Moreover, the probability that v is contained in a component of order $(2 + o(1))\varepsilon n$ is also $(2 + o(1))\varepsilon$.*

Our next result bounds from above the probability that a vertex v has many descendants in $D(n,p)$.

Lemma 6.4. *Let $np = 1 + \varepsilon$, where $|\varepsilon| \leq 1/2$ and suppose that $k \leq |\varepsilon|n/6$. Then, for some absolute constant c, we have*

$$\mathbb{P}[S(n,p) = k] \leq ck^{-3/2} \exp(-k\varepsilon^2/12) \leq ck^{-3/2}. \tag{6.1}$$

Proof. As we have already noted, the probability that $S(n,p)=k$ is equal to the probability that a vertex in $G(n,p)$ is contained in a component of order k. Let X_k be the number of components of order k in $G(n,p)$. Bollobás [8] (see also [11] p.132) showed that, for some absolute constant c',

$$\mathbb{E}[X_k] \le c'nk^{-5/2}\exp(-k\varepsilon^2/2+k\varepsilon^3/3+k^2\varepsilon/2n).$$

Since for $|\varepsilon|\le 1/2$ and $k\le|\varepsilon|n/6$ we have

$$-k\varepsilon^2/2+k\varepsilon^3/3+k^2\varepsilon/2n \le -k\varepsilon^2/12$$

and

$$\mathbb{P}[S(n,p)=k]=\frac{k}{n}\mathbb{E}[X_k]\,,$$

Equation (6.1) follows. □

From Lemmas 6.3 and 6.4 we get the following result.

Lemma 6.5. *Let $p=1+\varepsilon$, where $\varepsilon=\varepsilon(n)\to 0$, and $\bar{k}=o(n^{2/3})$. Then, for some absolute constant c,*

$$\mathbb{E}\Big[\Big|\overrightarrow{S}(v)\Big| \;\Big|\; \Big|\overrightarrow{S}(v)\Big|\le\bar{k}\Big] \le 2c\sqrt{\bar{k}}\,.$$

Proof. From Lemmas 6.3 and 6.4 we get

$$\begin{aligned}\mathbb{E}\Big[\Big|\overrightarrow{S}(v)\Big| \;\Big|\; \Big|\overrightarrow{S}(v)\Big|\le\bar{k}\Big] &\le \sum_{k=1}^{\bar{k}} k\frac{\mathbb{P}\Big[\Big|\overrightarrow{S}(v)\Big|=k\Big]}{\mathbb{P}\Big[\Big|\overrightarrow{S}(v)\Big|\le\bar{k}\Big]}\\ &\le \frac{1}{1-2\varepsilon+o(\varepsilon)}\sum_{k=1}^{\bar{k}} ck^{-1/2}\le 2c\sqrt{\bar{k}},\end{aligned}$$

so the assertion follows. □

6.2 The subcritical case

In this section we shall prove Theorem 6.1(i). In fact, we shall show the following slightly stronger result.

Theorem 6.6. *Let $np=1-\varepsilon$, where $\varepsilon=\varepsilon(n)\to 0$ but $n\varepsilon^3\to\infty$. Assume that a is a positive constant and let $X_s(n)$, $s\ge 1$, denote the order of the sth largest component of $D(n,p)$. Then a.a.s. $D(n,p)$ contains no complex components and*

$$\lim_{n\to\infty}\mathbb{P}(X_s<a/\varepsilon)=\sum_{i=0}^{s-1}\frac{\lambda_a^i}{i!}e^{-\lambda_a}\,,$$

where $\lambda_a=\int_a^\infty \frac{e^{-x}}{x}\,dx$.

Proof. Note that each complex component contains a pair of directed cycles C_1 and C_2 such that the intersection of the sets of their vertices spans a directed path (which, perhaps, consists of a single vertex). It is easy to see that there are at most $k^2k!$ such pairs on a given set of k vertices. Hence, the expected number of such pairs in $D(n,p)$, when $np = 1-\varepsilon$, is bounded from above by

$$\sum_{k=4}^{n} \binom{n}{k} k^2 k! p^{k+1} \le \frac{1}{n} \sum_{k=4}^{n} k^2 (1-\varepsilon)^k \le \frac{1}{n} \sum_{k=4}^{n} k^2 \exp(-\varepsilon k) = O\Big(\frac{1}{\varepsilon^3 n}\Big).$$

Thus, the probability that $D(n,p)$ contains a complex component is bounded from above by $O(1/n\varepsilon^3)$ and tends to 0 as $n \to \infty$. This completes the proof of the first part of Theorem 6.6.

Now let the random variable $Y_a(n)$ count the number of directed cycles larger than $m = \lceil a/\varepsilon \rceil$. Then

$$\mathbb{E}Y_a(n) = \sum_{k=m}^{n} \binom{n}{k} (k-1)! p^k = (1+o(1)) \sum_{k=m}^{n^{1/3}} \frac{1}{k} e^{-\varepsilon k} = (1+o(1)) \int_a^\infty \frac{e^{-x}}{x}\, dx = (1+o(1))\lambda_a\,.$$

One can easily check that for a given $r \ge 2$, the rth factorial moment of $Y_a(n)$ tends to λ_a^r as $n \to \infty$. Consequently, the random variable $Y_a(n)$ tends in distribution to a Poisson variable with expectation λ_a, and since clearly

$$\mathbb{P}(X_s(n) < a/\varepsilon) = \mathbb{P}(Y_a(n) \le s-1),$$

the assertion follows. □

6.3 The supercritical case

Throughout this section we study the structure of $D(n,p)$ in the supercritical phase when $np = 1+\varepsilon$, and $\varepsilon = \varepsilon(n) \to 0$ but $n\varepsilon^3 \to \infty$ as $n \to \infty$. Let us recall that $k_0 = 2\varepsilon^{-2} \log n\varepsilon^3$.

Let X denote the number of "large" vertices which have both more than $1.9\varepsilon n$ descendants and more than $1.9\varepsilon n$ ancestors. The main difficulty in the proof of Theorem 6.1(ii) is to estimate the number of large vertices, which is done in the following lemma, whose proof we postpone until the next section.

Lemma 6.7. *If $np = 1 + \varepsilon$, where $\varepsilon = \varepsilon(n) \to 0$ but $n\varepsilon^3 \to \infty$, then the expectation of X is given by $\mathbb{E}X = (4+o(1))\varepsilon^2 n$, while for the variance of X we have $\operatorname{Var} X = o((\mathbb{E}X)^2)$.*

In particular, a.a.s. $X = (4 + o(1))\varepsilon^2 n$.

Once we assume that Lemma 6.7 holds, the rest of the proof of Theorem 6.1(ii) is not so difficult. Let us start with the two following observations.

Lemma 6.8. *If $np = 1+\varepsilon$, where $\varepsilon = \varepsilon(n) \to 0$ but $n\varepsilon^3 \to \infty$, then a.a.s. at most one component of $D(n,p)$ contains cycles longer than $2\log\log n\varepsilon^3/\varepsilon$.*

Proof. Let v denote a large vertex of $D(n,p)$. (From Lemma 6.7 we know that a.a.s. at least one such vertex exists.) If we remove from $D(n,p)$ all vertices of $\overrightarrow{S}(v)$, we get a digraph with $n' = n - |\overrightarrow{S}(v)| \le n - 1.9\varepsilon n$ vertices. Since

$$n'p \le np - |\overrightarrow{S}(v)|p \le (1 - 0.8\varepsilon)n' ,$$

from Theorem 6.1(ii) we deduce that a.a.s. $D(n,p) \setminus \overrightarrow{S}(v)$ contains no cycles longer than, say, $\log\log n\varepsilon^3/\varepsilon$. An analogous argument shows that a.a.s. no such cycles are contained in $D(n,p)\setminus \overleftarrow{S}(v)$. Hence, a.a.s. all cycles longer that $2\log\log n\varepsilon^3/\varepsilon$ must belong to a strongly connected component containing v. □

Lemma 6.9. *If $np = 1 + \varepsilon$, where $\varepsilon = \varepsilon(n) \to 0$ but $n\varepsilon^3 \to \infty$, then a.a.s. only $o(\varepsilon^2 n)$ large vertices of $D(n,p)$ are not contained in cycles longer than $2\log\log n\varepsilon^3/\varepsilon$.*

Proof. Let v be a vertex of $D(n,p)$. Now generate $\overrightarrow{S}(v)$ with the breadth-first search process. The probability that $|\overrightarrow{S}(v)| \ge 1.8\varepsilon n$ is $(2 + o(1))\varepsilon$ (see Lemma 6.3). If v is large, then either $\overleftarrow{S}(v) \cap \overrightarrow{S}(v) = \{v\}$, or some in-neighbours of v are in $\overrightarrow{S}(v)$.

The probability that $\overleftarrow{S}(v) \cap \overrightarrow{S}(v) = \{v\}$ is bounded from above by the probability that some component of $G(n',p)$, $n'p = 1 - 0.8\varepsilon$, is larger than $1.9\varepsilon n'$, which is $o(\varepsilon)$ according to Lemma 6.3.

Now let $h = \log\log n\varepsilon^3/\varepsilon$. The expected number of vertices in $\overrightarrow{S_{\le h}}(v)$, conditioned on $\overrightarrow{S}(v)$ being large (i.e. the process does not become extinct quickly) is given by

$$\begin{aligned}\mathbb{E}[\overrightarrow{S_{\le h}}(v)|\overrightarrow{S_h}(v) > 0] &\le \frac{\mathbb{E}[\overrightarrow{S_{\le h}}(v)]}{\mathbb{P}[\overrightarrow{S_h}(v) > 0]} \le O\left(\sum_{i=1}^{h} \frac{(1+\varepsilon)^i}{2\varepsilon}\right) \\ &\le O\left(\frac{(1+\varepsilon)^h}{\varepsilon^2}\right) \le O\left(\frac{\log n\varepsilon^3}{\varepsilon^2}\right) .\end{aligned}$$

Consequently, the probability that there is an arc starting at a vertex in $\overrightarrow{S_{\le h}}(v)$ with the end at v is bounded from above by

$$O\left(\frac{\log n\varepsilon^3}{n\varepsilon^2}\right) = o(\varepsilon)\,.$$

Thus the probability that a vertex is large but either does not belong to a directed cycle or belongs to a cycle shorter than h, is bounded from above by

$$2\varepsilon(\varepsilon/\log\varepsilon^3 n + o(\varepsilon)) = o(\varepsilon^2)\,.$$

Now the assertion follows from Markov's inequality. □

Proof of Theorem 6.1(ii). From Lemmas 6.7 and 6.9 we infer that a.a.s. $(4+o(1))\varepsilon^2 n$ large vertices belong to components which contain cycles longer than $2\log\log n\varepsilon^3/\varepsilon$, while Lemma 6.8 ensures that such a component is unique.

Finally, observe that $D(n,p)$, with $np = 1+\varepsilon$, can be viewed as obtained from the digraph $D(n,p')$, where $np' = 1+\varepsilon/2$, by adding to it arcs of $D(n,p'')$, where $1-p = (1-p')(1-p'')$. Thus, a.a.s. the largest component L of $D(n,p)$, which has $(4+o(1))\varepsilon^2 n$ vertices, contains the largest component L' of $D(n,p')$, which contains $(1+o(1))\varepsilon^2 n$ vertices. But then, clearly, L cannot be a directed cycle, so it must be a complex component. □

6.4 Proof of main lemma

This section is devoted to the proof of Lemma 6.7. Let A denote the set of all large vertices of $D(n,p)$, that is the set of those vertices which have both more than $1.9\varepsilon n$ descendants and more than $1.9\varepsilon n$ ancestors. Instead of calculating the expectation and variance of $X = |A|$, we shall estimate the number Y of elements in a certain set B which overlaps with A, and then show that the symmetric difference between A and B is likely to be small.

Let $g = 2\varepsilon^{-1}\log\log n\varepsilon^3$, and let B denote the set of those vertices from which the breadth-first search lasts for at least g generations in both directions. In other words, B is the set of vertices v such that $\overrightarrow{S_g}(v) \neq \emptyset$ and $\overleftarrow{S_g}(v) \neq \emptyset$. We will show that a.a.s. $Y = |B| = (4+o(1))\varepsilon n$, and then show that a.a.s. $|X-Y| = o(\varepsilon^2 n)$.

Before calculating the expectation and variance of Y, we will state a number of inequalities which we shall use. Let v be a vertex in $D(n,p)$, and let D' be a set of vertices in $D(n,p)$ such that $|D'| = n' \le n$. Then $|\overrightarrow{S}(D',v)|$ and $|\overleftarrow{S}(D',v)|$ have the same distribution as the random variable $S(n',p)$,

while for every $d \geq 0$, $|\overrightarrow{S_d}(D', v)|$ and $|\overleftarrow{S_d}(D', v)|$ have the same distribution as $S_d(n', p)$.

Lemma 6.3 implies that

$$\mathbb{P}\left[S(n', p) \geq k_0\right] \leq (2 + o(1))\varepsilon.$$

If we moreover assume that $n' = n - o(\varepsilon n)$, then

$$\mathbb{P}\left[S(n', p) \geq k_0\right] = (2 + o(1))\varepsilon.$$

The expected number of neighbours of a vertex is $(n-1)p \leq 1+\varepsilon$, so the expected number of vertices in the ith generation of the breadth-first search starting at v is bounded from above by $(1+\varepsilon)^i$. Thus

$$\mathbb{E}\left[S_g(n', p)\right] \leq (1+\varepsilon)^g = O\left(\log^2 n\varepsilon^3\right),$$

while

$$\mathbb{E}\left[S_{\leq g}(n', p)\right] \leq \sum_{i=0}^{g}(1+\varepsilon)^i = O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right). \tag{6.2}$$

By Markov's inequality,

$$\mathbb{P}\left[S_{\leq g}(n', p) \geq \frac{\varepsilon n}{\log n\varepsilon^3}\right] \leq \frac{\mathbb{E}\left[S_{\leq g}(n', p)\right]}{\varepsilon n/\log n\varepsilon^3} = \frac{O\left(\log^3 n\varepsilon^3\right)}{n\varepsilon^2} = o(\varepsilon). \tag{6.3}$$

Since

$$\mathbb{P}\left[S(n', p) \geq \varepsilon n\right] \leq \mathbb{P}\left[S_{\leq g}(n', p) \geq \varepsilon n\right] + \mathbb{P}\left[S_g(n', p) > 0\right],$$

and $\mathbb{P}\left[S(n', p) \geq \varepsilon n\right] = (2 + o(1))\varepsilon$ if $n' = n - o(\varepsilon n)$ by Lemma 6.3, it follows from (6.3) that

$$\mathbb{P}\left[S_g(n', p) > 0\right] = (2 + o(1))\varepsilon, \tag{6.4}$$

provided that $n' = n - o(\varepsilon n)$. If Z is a nonnegative random variable and A is any event with $\mathbb{P}[A] > 0$, then

$$\mathbb{E}\left[Z|A\right] \leq \frac{\mathbb{E}\left[Z\right]}{\mathbb{P}\left[A\right]}. \tag{6.5}$$

If A is the event that $Z > 0$, then we have equality in (6.5). Conditioning on the breadth-first search from v lasting at least g generations, we get

$$\mathbb{E}\left[S_g(n', p)|S_g(n', p) > 0\right] = \frac{\mathbb{E}\left[S_g(n', p)\right]}{\mathbb{P}\left[S_g(n', p) > 0\right]} = O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right), \tag{6.6}$$

and

$$\mathbb{E}\left[S_{\leq g}(n',p)|S_g(n',p)>0\right] \leq \frac{\mathbb{E}\left[S_{\leq g}(n',p)\right]}{\mathbb{P}\left[S_g(n',p)>0\right]} = O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon^2}\right). \tag{6.7}$$

If we instead condition on the process dying out before g generations, we obtain

$$\begin{aligned}\mathbb{E}\left[S(n',p)|S_g(n',p)=0\right] &= \mathbb{E}\left[S_{\leq g}(n',p)|S_g(n',p)=0\right] \leq \frac{\mathbb{E}\left[S_{\leq g}(n',p)\right]}{\mathbb{P}\left[S_g(n',p)=0\right]} \\ &= O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right). \end{aligned} \tag{6.8}$$

Finally, from Lemma 6.5 it follows that

$$\mathbb{E}\left[S(n',p)|S(n',p)<k_0\right] = O\left(\frac{\log^{1/2} n\varepsilon^3}{\varepsilon}\right). \tag{6.9}$$

First we will determine the expected number of vertices in B. Let v be a vertex in $D(n,p)$. Using the breadth-first search we will first expose the set

$$\overrightarrow{V} := \overrightarrow{S_{\leq g}}(v),$$

and then the set

$$\overleftarrow{V} := \overleftarrow{S_{\leq g}}(D \setminus \overrightarrow{V}, v).$$

Let $V = \overrightarrow{V} \cup \overleftarrow{V}$. Note that in the process of exposing the sets $\overrightarrow{V}$ and $\overleftarrow{V}$ no potential arc is tested more than once.

We moreover let $\overrightarrow{V_g} = \overrightarrow{S_g}(v)$ and $\overleftarrow{V_g} = \overleftarrow{S_g}(v)$. The significance of these sets is that the vertices in $\overrightarrow{V_g}$ are the only vertices in $\overrightarrow{V}$ from which there can be an arc to $D \setminus \overrightarrow{V}$. Similarly, the vertices in $\overleftarrow{V_g}$ are the only vertices in $\overleftarrow{V}$ to which there can be an arc from $D \setminus V$.

Lemma 6.10. $\mathbb{E}[Y] = (4+o(1))\varepsilon^2 n$.

Proof. Let v be a vertex, and let $\overrightarrow{V}$ and $\overleftarrow{V}$ be defined as above. From (6.4) it follows that the probability that $\overrightarrow{V_g} \neq \emptyset$ is $(2+o(1))\varepsilon$. If, in addition, $|\overrightarrow{V}| = o(\varepsilon n)$, the probability that $\overleftarrow{V_g} \neq \emptyset$ is also $(2+o(1))\varepsilon$. The probability that $|\overrightarrow{V}| \geq \frac{\varepsilon n}{\log n\varepsilon^3}$ is $o(\varepsilon)$ by (6.3), so it follows that the probability that both $\overrightarrow{V_g} \neq \emptyset$ and $\overleftarrow{V_g} \neq \emptyset$ is $(4+o(1))\varepsilon^2 n$.

The set $\overleftarrow{V}$ may be strictly smaller than $\overleftarrow{S}(v)$, so it is possible that $\overleftarrow{S_g}(v) \neq \emptyset$ although $\overleftarrow{V_g} = \emptyset$. This may happen only if there are arcs from

$\overrightarrow{V}$ to $\overleftarrow{V}$. As noted above there can be no arc from $\overrightarrow{V_{<g}}$ to $D \setminus \overrightarrow{V}$, so we only have to consider arcs going from $\overrightarrow{V_g}$ to $\overleftarrow{V}$. The expected number of such arcs is $\mathbb{E}\left[|\overrightarrow{V_g}| \cdot |\overleftarrow{V}|\right] p$. The random variables $|\overrightarrow{V_g}|$ and $|\overleftarrow{V}|$ are not independent, but we note that the bound for $|\overleftarrow{V}|$ given by (6.8) holds regardless of the size of $\overrightarrow{V_g}$. Hence, the probability that there is an arc from $\overrightarrow{V_g}$ to $\overleftarrow{V}$ is bounded from above by

$$O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right) O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right) p = O\left(\frac{\log^4 n\varepsilon^3}{n\varepsilon^2}\right) = o(\varepsilon)$$

according to (6.6) and (6.8). This completes the proof of the lemma. □

Lemma 6.11. *Let v be a vertex of $D(n,p)$, and let $\mathcal{C}_1$ denote the event that $\overrightarrow{S_g}(v) \neq \emptyset$ and $\mathcal{C}_2$ denote the event that $|\overrightarrow{S}(v)| \geq k_0$. Then,*

$$\mathbb{P}\left[(\mathcal{C}_1 \wedge \overline{\mathcal{C}_2}) \vee (\overline{\mathcal{C}_1} \wedge \mathcal{C}_2)\right] = o(\varepsilon).$$

Proof. Suppose first that $\overrightarrow{S_g}(v) \neq \emptyset$. Thus the breadth-first search starting at v lasts for at least g generations. Remembering the duality of the models $D(n,p)$ and $G(n,p)$, the probability that this happens in $D(n,p)$ is equal to the probability that it happens in $G(n,p)$. In the model $G(n,p)$ this can only happen if v is in a component of diameter at least g. Let $Z_{n,p}(g)$ be the number of components in $G(n,p)$ with fewer than k_0 vertices and diameter at least g. Łuczak [37] showed that

$$\mathbb{E}\left[Z_{n,p}(g)\right] = (2 + o(1))n\varepsilon^3(1-\varepsilon)^g.$$

The probability that the breadth-first search starting at v lasts at least g generations but does not generate k_0 vertices is therefore bounded from above by

$$\frac{k_0 \mathbb{E}\left[Z_{n,p}(g)\right]}{n} = O\left(\frac{\varepsilon}{\log n\varepsilon^3}\right) = o(\varepsilon),$$

and hence

$$\mathbb{P}\left[\overrightarrow{S_g}(v) \neq \emptyset \wedge |\overrightarrow{S}(v)| < k_0\right] = o(\varepsilon).$$

On the other hand, suppose that $\overrightarrow{S_g}(v) = \emptyset$. By (6.3) the probability that $S_{\leq g}(n', p) \geq \varepsilon n / \log n\varepsilon^3$ is $o(\varepsilon)$ if $\overrightarrow{S_g}(v) = \emptyset$, whereas the probability that $k_0 \leq |\overrightarrow{S}(v)| \leq \varepsilon n / \log n\varepsilon^3$ is $o(\varepsilon)$ by Lemma 6.3. Thus

$$\mathbb{P}\left[\overrightarrow{S_g}(v) = \emptyset \wedge |\overrightarrow{S}(v)| \geq k_0\right] = o(\varepsilon),$$

which finishes the proof. □

The next step in the proof is to bound the variance, in order that we may conclude that Y is likely to be close to its expectation.

Lemma 6.12. $\operatorname{Var} Y = o(\mathbb{E}[Y]^2)$.

Proof. We will prove this lemma by showing that $\mathbb{E}[Y(Y-1)] = (1 + o(1))\mathbb{E}[Y]^2$. That is, we will calculate the number of pairs of vertices v and w such that we have $\overrightarrow{S_g}(v) \neq \emptyset$, $\overleftarrow{S_g}(v) \neq \emptyset$, $\overrightarrow{S_g}(w) \neq \emptyset$ and $\overleftarrow{S_g}(w) \neq \emptyset$.

We assume that v is a vertex such that $\overrightarrow{S_g}(v) \neq \emptyset$ and $\overleftarrow{S_g}(v) \neq \emptyset$, and count the number of vertices $w \neq v$ such that $\overrightarrow{S_g}(w) \neq \emptyset$ and $\overleftarrow{S_g}(w) \neq \emptyset$. Our goal is to show that the expected number of such vertices is $(4 + o(1))\varepsilon^2 n$. As previously, we first expose the descendants and ancestors of v, producing the sets $\overrightarrow{V}$ and $\overleftarrow{V}$. We then expose the descendants and ancestors of w in a similar way, letting

$$\overrightarrow{W} = \overrightarrow{S_{\leq g}}(D \setminus V, w)$$

and

$$\overleftarrow{W} = \overleftarrow{S_{\leq g}}(D \setminus (V \cup \overrightarrow{W}), w).$$

We also define the sets $\overrightarrow{W_g} = \overrightarrow{S_g}(D \setminus V, w)$ and $\overleftarrow{W_g} = \overleftarrow{S_g}(D \setminus (V \cup \overrightarrow{W}), w)$. Finally let $W = \overrightarrow{W} \cup \overleftarrow{W}$ and $D' = D \setminus (V \cup W)$.

We will first show that

$$\mathbb{P}\left[\overrightarrow{S_g}(w) \neq \emptyset \middle| v \in B \wedge w \notin V\right] = (2 + o(1))\varepsilon. \tag{6.10}$$

The probability that $\overrightarrow{W_g} \neq \emptyset$ is $(2 + o(1))\varepsilon$ by (6.4). In order to show that (6.10) holds, we will therefore show that if $\overrightarrow{W_g} = \emptyset$, then the probability that $\overrightarrow{S_g}(w) \neq \emptyset$ is $o(\varepsilon)$. In other words, if $\mathcal{A}_1$ is the event that $v \in B$, $w \notin V$ and $\overrightarrow{W_g} = \emptyset$, we want to show that

$$\mathbb{P}\left[\overrightarrow{S_g}(w) \neq \emptyset \middle| \mathcal{A}_1\right] = o(\varepsilon). \tag{6.11}$$

If $\mathcal{A}_1$ holds and $\overrightarrow{S_g}(w) \neq \emptyset$, it must be because there are arcs from $\overrightarrow{W}$ to V. Note that there can be no arcs from $\overrightarrow{W}$ to $\overleftarrow{V_{<g}}$, since the origin of any such arc would be in $\overleftarrow{V}$. We therefore only have to consider arcs from $\overrightarrow{W}$ to $\overleftarrow{V_g}$ and $\overrightarrow{V}$. We will show below that the probability that there are arcs to $\overleftarrow{V_g}$ is $o(\varepsilon)$. The probability that there are arcs to $\overrightarrow{V}$ is significantly greater, so we will have to look more carefully at this situation.

We will distinguish between two types of vertices in $\overrightarrow{V}$, which we shall call *malevolent* and *benevolent* vertices. Loosely speaking, the malevolent vertices

are those which have many descendants, while the benevolent vertices have few descendants. We will then show that w has a malevolent descendant with probability $o(\varepsilon)$. On the other hand, w may have several benevolent descendants, but we will show that it is unlikely to have so many of them that the total number of descendants of w becomes k_0 or more. Because of Lemma 6.11, this will be sufficient.

Recall that the set $\overrightarrow{V}$ was found using a breadth-first search starting with the vertex v. In the course of exposing the vertices in $\overrightarrow{V}$ we have therefore exposed several arcs between vertices in $\overrightarrow{V}$, such that we have a directed tree, rooted at v. Let us call this tree $\overrightarrow{T}$. This tree has the same vertex set as $\overrightarrow{V}$, while the set of arcs of $\overrightarrow{T}$ is a subset of the set of arcs of $D(n,p)$ joining vertices in $\overrightarrow{V}$.

If x is a vertex in $\overrightarrow{V}$, we define $\overrightarrow{T_x}$ to be the subtree of $\overrightarrow{T}$ rooted at x. In precise terms, $\overrightarrow{T_x}$ consists of those vertices $y \in \overrightarrow{T}$ such that the unique path from v to y in $\overrightarrow{T}$ passes through x. Let $\overrightarrow{T_x}' = \overrightarrow{T_x} \cap \overrightarrow{V_g}$ be the set of descendants of x in $\overrightarrow{T}$ which are in the last generation of $\overrightarrow{V}$. These are the only vertices in $\overrightarrow{T_x}$ from which there can be an arc to $D \setminus \overrightarrow{V}$. Note that this set may be empty. We moreover define the set $\overrightarrow{U_x}$ by

$$\overrightarrow{U_x} = \overrightarrow{T_x} \cup \bigcup_{y \in \overrightarrow{T_x}'} \overrightarrow{S_{\le g}}(D', y).$$

Thus, to obtain $\overrightarrow{U_x}$, we first find the subtree of $\overrightarrow{T}$ rooted at x, and then continue the breadth-first search for up to g generations in D'. We say that x is *malevolent* if $\overrightarrow{T_x}'$ is nonempty and $S_g(D', y) \neq \emptyset$ for some $y \in T_x'$. Otherwise we say that x is *benevolent*.

We will now attempt to find the vertices in $\overrightarrow{V}$ which are descendants of w. We first expose all arcs from $\overrightarrow{W}$ to $\overrightarrow{V}$ and let $\overrightarrow{Z_1}$ be the set of vertices in $\overrightarrow{V}$ to which there is an arc with origin in $\overrightarrow{W}$. Then we expose the descendants of these vertices, but without exposing any more internal arcs in $\overrightarrow{V}$, letting

$$\overrightarrow{U_1} = \bigcup_{x \in \overrightarrow{Z_1}} \overrightarrow{U_x}.$$

Having found these extra descendants of w, we expose the possible arcs going from $\overrightarrow{U_1}$ to $\overrightarrow{V}$, and continue as above. We define the sets $\overrightarrow{Z_2}, \overrightarrow{Z_3}, \dots$ and $\overrightarrow{U_2}, \overrightarrow{U_3}, \dots$ as follows: For $m \ge 2$, let $\overrightarrow{Z_m}$ be the set of vertices in $\overrightarrow{V} \setminus \left(\overrightarrow{U_1} \cup \dots \cup \overrightarrow{U_{m-1}}\right)$ to which there is an arc from $\overrightarrow{U_{m-1}}$, and let

$$\overrightarrow{U_m} = \bigcup_{x \in \overrightarrow{Z_m}} \overrightarrow{U_x} \setminus \bigcup_{i=1}^{m-1} \overrightarrow{U_i}.$$

Finally, let $\overrightarrow{U} = \bigcup_{m\geq 1} \overrightarrow{U_m}$ be the set of all descendants of w found by this procedure. Observe that by this process we are guaranteed to expose the entire set of $\overrightarrow{U}$, even though $\overrightarrow{U_x}$ and $\overrightarrow{U_y}$ may overlap for many choices of x and y. Further observe that if none of the vertices in $\overrightarrow{U}$ are malevolent, then the breadth-first searches used to expose the sets $\overrightarrow{U_1}, \overrightarrow{U_2}, \ldots$ are allowed to continue until no more descendants can be found in D'. In this case there are therefore no arcs from $\overrightarrow{U}$ to the rest of D'. Thus if there is also no arc from $\overrightarrow{W} \cup \overrightarrow{U}$ to $\overleftarrow{V}$ or $\overleftarrow{W}$, then $\overrightarrow{W} \cup \overrightarrow{U}$ comprises all descendants of w. We can therefore conclude that if $\overrightarrow{S_g}(w) \neq \emptyset$ and $\mathcal{A}_1$ holds, at least one of the following events must happen.

(i) w has a malevolent descendant.

(ii) There is an arc from $\overrightarrow{W}$ or $\overrightarrow{U}$ to $\overleftarrow{V}$ or $\overleftarrow{W}$.

(iii) $|\overrightarrow{W} \cup \overrightarrow{U}| \geq k_0$.

(iv) $\overrightarrow{S_g}(w) \neq \emptyset$ and $|\overrightarrow{S}(w)| < k_0$.

We will show that each of these events happens with probability $o(\varepsilon)$. That (iv) holds with probability $o(\varepsilon)$ follows from Lemma 6.11, so we only have to consider the three first events. From (6.6) and (6.9) we have

$$\mathbb{E}\left[|\overrightarrow{V_g}|\middle|\mathcal{A}_1\right] = O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right) \tag{6.12}$$

and

$$\mathbb{E}\left[|\overrightarrow{W}|\middle|\mathcal{A}_1\right] = O\left(\frac{\log^{1/2} n\varepsilon^3}{\varepsilon}\right). \tag{6.13}$$

Let us first count the expected number of malevolent vertices. If $y \in \overrightarrow{V_g}$, the probability that y is malevolent is simply the probability that $\overrightarrow{S_g}(D', y) \neq \emptyset$, which is bounded from above by $(2+o(1))\varepsilon$ according to (6.4). Every vertex in $\overrightarrow{V_g}$ has g ancestors in $\overrightarrow{T}$, so by (6.12) the expected number of malevolent vertices in $\overrightarrow{V}$ is bounded from above by

$$O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right)(2+o(1))\varepsilon g \leq O\left(\frac{\log^3 n\varepsilon^3}{\varepsilon}\right). \tag{6.14}$$

Now let us count the expected number of vertices in $\overrightarrow{U}_1, \overrightarrow{U}_2, \ldots$. Suppose that x is a randomly chosen vertex in $\overrightarrow{V}$ and let y be a vertex in $\overrightarrow{V_g}$. Since y

has g ancestors in $\overrightarrow{T}$, the probability that x is an ancestor of y in $\overrightarrow{T}$ is $g/|\overrightarrow{V}|$. The expected number of descendants of x in $\overrightarrow{T_x}'$ is therefore

$$\mathbb{E}\left[|\overrightarrow{T_x}'|\middle|\mathcal{A}_1\right] = \frac{g\mathbb{E}\left[|\overrightarrow{V_g}|\middle|\mathcal{A}_1\right]}{|\overrightarrow{V}|} = \frac{O\left(\varepsilon^{-2}\log^3 n\varepsilon^3\right)}{|\overrightarrow{V}|}. \tag{6.15}$$

There are at most g vertices in the unique path in $\overrightarrow{T_x}$ from x to y for every $y \in \overrightarrow{T_x}'$, so

$$\mathbb{E}\left[|\overrightarrow{T_x}|\middle|\mathcal{A}_1\right] \le \mathbb{E}\left[|\overrightarrow{T_x}'|\middle|\mathcal{A}_1\right] g = \frac{O\left(\varepsilon^{-3}\log^4 n\varepsilon^3\right)}{|\overrightarrow{V}|}. \tag{6.16}$$

If x is benevolent, then $\overrightarrow{S_g}(D', y) = \emptyset$ for every $y \in \overrightarrow{T_x}'$. Thus, by (6.8),

$$\begin{aligned}
\mathbb{E}\left[|\overrightarrow{U_x}|\middle|\mathcal{A}_1\right] &\le \mathbb{E}\left[|\overrightarrow{T_x}|\middle|\mathcal{A}_1\right] + \mathbb{E}\left[|\overrightarrow{T_x}'|\middle|\mathcal{A}_1\right] O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right) \\
&= \frac{O\left(\varepsilon^{-3}\log^4 n\varepsilon^3\right)}{|\overrightarrow{V}|} + \frac{O\left(\varepsilon^{-2}\log^3 n\varepsilon^3\right)}{|\overrightarrow{V}|} O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right) \\
&= \frac{O\left(\varepsilon^{-3}\log^5 n\varepsilon^3\right)}{|\overrightarrow{V}|}.
\end{aligned} \tag{6.17}$$

The expected number of arcs going from $\overrightarrow{W}$ to $\overrightarrow{V}$ is $\mathbb{E}\left[|\overrightarrow{W}| \cdot |\overrightarrow{V}|\middle|\mathcal{A}_1\right] p$. Note that (6.13) holds even if we condition on the size of $\overrightarrow{V}$. Thus

$$\mathbb{E}\left[|\overrightarrow{Z_1}|\middle|\mathcal{A}_1\right] \le O\left(\frac{\log^{1/2} n\varepsilon^3}{n\varepsilon}\right)|\overrightarrow{V}|. \tag{6.18}$$

For $m \ge 1$, let $\mathcal{B}_m$ be the event that all the sets $\overrightarrow{U_1}, \ldots, \overrightarrow{U_m}$ contain only benevolent vertices. Then, from (6.17) and (6.18),

$$\begin{aligned}
\mathbb{E}\left[|\overrightarrow{U_1}|\middle|\mathcal{A}_1 \wedge \mathcal{B}_1\right] &\le \mathbb{E}\left[|\overrightarrow{Z_1}|\middle|\mathcal{A}_1\right] \mathbb{E}\left[U_x|\mathcal{A}_1\right] \\
&= O\left(\frac{\log^{11/2} n\varepsilon^3}{n\varepsilon^4}\right).
\end{aligned} \tag{6.19}$$

Generally, if $\mathcal{B}_{m+1}$ holds, then

$$\mathbb{E}\left[|\overrightarrow{Z_{m+1}}|\middle|\mathcal{A}_1 \wedge \mathcal{B}_{m+1}\right] \le \mathbb{E}\left[|\overrightarrow{U_m}| \cdot |\overrightarrow{V}|\middle|\mathcal{A}_1 \wedge \mathcal{B}_m\right] p. \tag{6.20}$$

Therefore, according to (6.17),

$$\begin{aligned}\mathbb{E}\left[|\overrightarrow{U_{m+1}}|\middle|\mathcal{A}_1 \wedge \mathcal{B}_{m+1}\right] &\leq \mathbb{E}\left[|\overrightarrow{Z_{m+1}}|\middle|\mathcal{A}_1 \wedge \mathcal{B}_{m+1}\right]\mathbb{E}\left[|\overrightarrow{U_x}|\middle|x \text{ is benevolent}\right] \\ &= \mathbb{E}\left[|\overrightarrow{U_m}|\middle|\mathcal{A}_1 \wedge \mathcal{B}_m\right] O\left(\frac{\log^5 n\varepsilon^3}{n\varepsilon^3}\right). \qquad (6.21)\end{aligned}$$

By induction it follows that

$$\mathbb{E}\left[|\overrightarrow{U_m}|\middle|\mathcal{A}_1 \wedge B_m\right] \leq O\left(\varepsilon^{-1}C^m\frac{\log^{5m+1/2} n\varepsilon^3}{(n\varepsilon^3)^m}\right) \qquad (6.22)$$

for $m \geq 1$, where C is the constant implicit in (6.21).

We can now calculate the probability that w has a malevolent descendant. The expected number of arcs from $\overrightarrow{W}$ to a malevolent vertex in $\overrightarrow{V}$ is, according to (6.13) and (6.14),

$$O\left(\frac{\log^{1/2} n\varepsilon^3}{\varepsilon}\right) O\left(\frac{\log^3 n\varepsilon^3}{\varepsilon}\right) p = O\left(\frac{\log^{7/2} n\varepsilon^3}{n\varepsilon^2}\right).$$

If we assume that $\mathcal{B}_m$ holds, then the expected number of arcs from $\overrightarrow{U_m}$ to a malevolent vertex is bounded from above by

$$O\left(\varepsilon^{-1}C^m\frac{\log^{5m+1/2} n\varepsilon^3}{(n\varepsilon^3)^m}\right) O\left(\frac{\log^3 n\varepsilon^3}{\varepsilon}\right) p = O\left(\varepsilon C^m\frac{\log^{5m+7/2} n\varepsilon^3}{(n\varepsilon^3)^{m+1}}\right).$$

Hence, the probability that $\overrightarrow{U}$ contains a malevolent vertex is

$$O\left(\varepsilon\sum_{m\geq 0}\frac{C^m \log^{5m+7/2} n\varepsilon^3}{(n\varepsilon^3)^{m+1}}\right) = O\left(\frac{\log^{7/2} n\varepsilon^3}{n\varepsilon^2}\right) = o(\varepsilon).$$

This proves that the event (i) happens with probability $o(\varepsilon)$.

Now we consider the cases (ii) and (iii). From (6.22) we find that

$$\mathbb{E}\left[|\overrightarrow{U}|\middle|\mathcal{A}_1\right] = O\left(\varepsilon^{-1}\sum_{m\geq 1} C^m\frac{\log^{5m+1/2} n\varepsilon^3}{(n\varepsilon^3)^m}\right) = O\left(\frac{\log^{11/2} n\varepsilon^3}{n\varepsilon^4}\right), \qquad (6.23)$$

so by (6.13) and (6.23),

$$\mathbb{E}\left[|\overrightarrow{W}| + |\overrightarrow{U}|\middle|\mathcal{A}_1\right] = O\left(\frac{\log^{1/2} n\varepsilon^3}{\varepsilon}\right). \qquad (6.24)$$

Markov's inequality then gives us

$$\mathbb{P}\left[|\overrightarrow{W}| + |\overrightarrow{U}| \geq k_0\right] \leq \frac{\mathbb{E}\left[|\overrightarrow{W}| + |\overrightarrow{U}|\right]}{k_0} = \frac{O\left(\varepsilon^{-1}\log^{1/2} n\varepsilon^3\right)}{2\varepsilon^{-2}\log n\varepsilon^3} = o(\varepsilon),$$

thereby proving that (iii) happens with probability $o(\varepsilon)$.

What remains is to calculate the probability that there is an arc from $\overrightarrow{W} \cup \overrightarrow{U}$ to $\overleftarrow{V}$ or $\overleftarrow{W}$. We first note that there can be no arcs from $\overrightarrow{W} \cup \overrightarrow{U}$ to $\overleftarrow{V_{<g}}$, except from vertices in $\overrightarrow{V_g}$. The equation (6.12) also holds for $\overleftarrow{V_g}$, so together with (6.24) this shows that the probability that there is an arc from $\overrightarrow{W} \cup \overrightarrow{U}$ to $\overleftarrow{V_g}$ is

$$O\left(\frac{\log^{1/2} n\varepsilon^3}{\varepsilon}\right) O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right) p = O\left(\frac{\log^{5/2} n\varepsilon^3}{n\varepsilon^2}\right) = o(\varepsilon).$$

We have made no assumptions regarding the order of $\overleftarrow{W}$, so the expected size is given by (6.2). The probability that there is an arc from $\overrightarrow{W} \cup \overrightarrow{U}$ to $\overleftarrow{W}$ is therefore

$$O\left(\frac{\log^{1/2} n\varepsilon^3}{\varepsilon}\right) O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right) p = O\left(\frac{\log^{5/2} n\varepsilon^3}{n\varepsilon^2}\right) = o(\varepsilon).$$

The only arcs we have yet to consider are therefore the ones from $\overrightarrow{U} \cap \overrightarrow{V_g}$ to $\overleftarrow{V_{<g}}$. Let $\overrightarrow{Z} = \bigcup_{m\geq 1} \overrightarrow{Z_m}$. From (6.18), (6.20) and (6.22) it follows that $\mathbb{E}[\overrightarrow{Z}] = O\left(\frac{\log^{1/2} n\varepsilon^3}{n\varepsilon}\right)|\overrightarrow{V}|$. By (6.15) it then follows that

$$\mathbb{E}\left[|\overrightarrow{U} \cap \overrightarrow{V_g}|\middle|\mathcal{A}_1\right] = O\left(\frac{\log^{7/2} n\varepsilon^3}{n\varepsilon^3}\right).$$

The expected size of $\overleftarrow{V_{<g}}$ is bounded by (6.7), so the probability that there is an arc from $\overrightarrow{U} \cap \overrightarrow{V_g}$ to $\overleftarrow{V_{<g}}$ is bounded by

$$O\left(\frac{\log^{7/2} n\varepsilon^3}{n\varepsilon^3}\right) O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon^2}\right) p = o(\varepsilon).$$

This completes the proof of (6.11).

We now assume that w is a vertex such that $\overrightarrow{S_g}(w) \neq \emptyset$, and we want to calculate the probability that $\overleftarrow{S_g}(w) \neq \emptyset$. We will show that

$$\mathbb{P}\left[\overleftarrow{S_g}(w) \neq \emptyset\middle| v \in B \wedge w \notin V \wedge \overrightarrow{S_g}(w) \neq \emptyset\right] = (2 + o(1))\varepsilon. \tag{6.25}$$

The probability that $\overleftarrow{W_g} \neq \emptyset$ is $(2+o(1))\varepsilon$ by (6.4). Let $\mathcal{A}_2$ be the event that $v \in B$, $w \notin V$, $\overrightarrow{S_g}(w) \neq \emptyset$ and $\overleftarrow{W_g} = \emptyset$. In order to show that (6.25) holds, we will show that

$$\mathbb{P}\left[\overleftarrow{S_g}(w) \neq \emptyset | \mathcal{A}_2\right] = o(\varepsilon). \tag{6.26}$$

We want to avoid exposing arcs leading to vertices which we have exposed before, in order to avoid dependencies. In the previous step we may have had to expose some of the vertices in D' in order to determine whether $\overrightarrow{S_g}(w) \neq \emptyset$. To be precise, we may have exposed some of the vertices in $\overrightarrow{S_{\leq 2g}}(v) \setminus \overrightarrow{V}$, but we do not know how many. In order to keep our calculations as tidy as possible, we will simply assume that we have exposed *all* of the vertices in $\overrightarrow{S_{\leq 2g}}(D' \cup \overrightarrow{V}, v)$, and we call this set $\overrightarrow{V}'$. In addition we let $\overrightarrow{V_{2g}}' = \overrightarrow{S_{2g}}(D' \cup \overrightarrow{V}, v)$. We have

$$\mathbb{E}\left[|\overrightarrow{V}'| \middle| \overrightarrow{S_g}(v) \neq \emptyset\right] \leq \frac{\mathbb{E}\left[\sum_{i=0}^{2g}(1+\varepsilon)^i\right]}{\mathbb{P}\left[\overrightarrow{S_g}(v) \neq \emptyset\right]} = O\left(\frac{\log^4 n\varepsilon^3}{\varepsilon^2}\right) \tag{6.27}$$

and similarly

$$\mathbb{E}\left[|\overrightarrow{V_{2g}}'| \middle| \overrightarrow{S_g}(v) \neq \emptyset\right] = O\left(\frac{\log^4 n\varepsilon^3}{\varepsilon}\right). \tag{6.28}$$

We moreover let $D'' = D' \setminus V'$.

The calculations needed to prove (6.26) are very similar to in the previous case, and we will use the same notation as earlier, with the arrows reversed: in the course of exposing $\overleftarrow{V}$, we have exposed a directed tree, rooted at v, which we will call $\overleftarrow{T}$. If x is a vertex in $\overleftarrow{V}$, we define $\overleftarrow{T_x}$ to be the subtree of $\overleftarrow{T}$ rooted at x, and we let $\overleftarrow{T_x}' = \overleftarrow{T_x} \cap \overleftarrow{V_g}$. Furthermore $\overleftarrow{U_x}$ is defined such that

$$\overleftarrow{U_x} = \overleftarrow{T_x} \cup \bigcup_{y \in \overleftarrow{T_x}'} \overleftarrow{S_{\leq g}}(D'', v).$$

A vertex $y \in \overleftarrow{V_g}$ is malevolent if $\overleftarrow{S_g}(D'', y) \neq \emptyset$. A vertex $x \in \overleftarrow{V_{<g}}$ is malevolent if it has a malevolent ancestor in the tree $\overleftarrow{T}$. Vertices in $\overleftarrow{V}$ which are not malevolent are benevolent.

We let $\overleftarrow{Z_1}$ be the set of vertices x in $\overleftarrow{V}$ such that there is an arc from x to some vertex in $\overleftarrow{W}$, and we let

$$\overleftarrow{U_1} = \bigcup_{x \in \overleftarrow{Z_1}} \overleftarrow{U_x}.$$

For $m \geq 1$, $\overleftarrow{Z_{m+1}}$ is the set of vertices x in $\overleftarrow{V} \setminus (\overleftarrow{U_1} \cup \ldots \cup \overleftarrow{U_m})$, such that there is an arc from x to some vertex in $\overleftarrow{U_m}$, and for $m \geq 2$,

$$\overleftarrow{U_m} = \bigcup_{x \in \overleftarrow{Z_m}} \overleftarrow{U_x} \setminus \bigcup_{i=1}^{m-1} \overleftarrow{U_i}.$$

Finally $\overleftarrow{Z} = \bigcup_{m \geq 1} \overleftarrow{Z_m}$ and $\overleftarrow{U} = \bigcup_{m \geq 1} \overleftarrow{U_m}$. As in the previous case, we see that if $\mathcal{A}_2$ is satisfied and $\overleftarrow{S_g}(w) \neq \emptyset$, then at least one of the following events must happen.

(i) w has a malevolent ancestor.

(ii) There is an arc from $\overrightarrow{V}'$ or $\overrightarrow{W}$ to $\overleftarrow{W}$ or $\overleftarrow{U}$.

(iii) $|\overleftarrow{W} \cup \overleftarrow{U}| \geq k_0$.

(iv) $\overleftarrow{S_g}(w) \neq \emptyset$ and $|\overleftarrow{S_{\leq g}}(w)| < k_0$.

Most of the calculations which show that each of these events happens with probability $o(\varepsilon)$ are largely identical to the previous case, so we will skip them. For example, the probability that there is an arc from $\overrightarrow{V_g}$ to $\overleftarrow{W}$ is virtually the same as the probability that, in the previous case, there was an arc from $\overrightarrow{W}$ to $\overleftarrow{V_g}$. The only differences from the previous case is that we have $\overrightarrow{W_g} \neq \emptyset$ by assumption, and we have to consider the possibility of arcs from $\overrightarrow{V_{2g}}'$ to $\overleftarrow{U}$ and from $\overrightarrow{V}'$ to $\overleftarrow{U} \cap \overleftarrow{V_g}$.

By similar calculations as earlier, we obtain

$$\mathbb{E}\left[\overrightarrow{W_g} \middle| \mathcal{A}_2\right] = O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right), \tag{6.29}$$

$$\mathbb{E}\left[\overleftarrow{W} \middle| \mathcal{A}_2\right] = O\left(\frac{\log^{1/2} n\varepsilon^3}{\varepsilon}\right), \tag{6.30}$$

$$\mathbb{E}\left[\overleftarrow{U} \middle| \mathcal{A}_2\right] = O\left(\frac{\log^{11/2} n\varepsilon^3}{n\varepsilon^4}\right), \tag{6.31}$$

and

$$\mathbb{E}\left[|\overleftarrow{U} \cap \overleftarrow{V_g}| \middle| \mathcal{A}_2\right] = O\left(\frac{\log^{7/2} n\varepsilon^3}{n\varepsilon^3}\right). \tag{6.32}$$

Thus, by (6.29), (6.30) and (6.31), the probability that there is an arc from $\overrightarrow{W_g}$ to $\overleftarrow{W}$ or $\overleftarrow{U}$ is

$$O\left(\frac{\log^2 n\varepsilon^3}{\varepsilon}\right) O\left(\frac{\log^{1/2} n\varepsilon^3}{\varepsilon}\right) p = O\left(\frac{\log^{5/2} n\varepsilon^3}{n\varepsilon^2}\right) = o(\varepsilon).$$

By (6.28) and (6.31), the probability that there is an arc from $\overrightarrow{V_{2g}}'$ to $\overleftarrow{U}$ is

$$O\left(\frac{\log^4 n\varepsilon^3}{\varepsilon}\right) O\left(\frac{\log^{11/2} n\varepsilon^3}{n\varepsilon^4}\right) p = O\left(\frac{\log^{19/2} n\varepsilon^3}{n^2\varepsilon^5}\right) = o(\varepsilon).$$

By (6.27), (6.7) and (6.31), the probability that there is an arc from $\overrightarrow{V}'$ or $\overrightarrow{W}$ to $\overleftarrow{U} \cap \overleftarrow{V_g}$ is

$$\left(O\left(\frac{\log^4 n\varepsilon^3}{\varepsilon^2}\right) + O\left(\frac{\log^{1/2} n\varepsilon^3}{\varepsilon^2}\right)\right) O\left(\frac{\log^{7/2} n\varepsilon^3}{n\varepsilon^3}\right) p = o(\varepsilon).$$

This implies that (6.26) holds. It now follows from (6.10) and (6.25) that if $v \in B$, then the expected number of vertices $w \notin V$ such that $w \in B$ is $(4 + o(1))\varepsilon^2 n$.

It remains to calculate the expected number of vertices $w \in V$ such that $w \in B$. Let us first count the number of such vertices in $\overrightarrow{V}$. Since by assumption $\overleftarrow{S_g}(v) \neq \emptyset$, for every vertex w in $\overrightarrow{V}$ we have $\overleftarrow{S_g}(w) \neq \emptyset$, so we have to bound the probability that $\overrightarrow{S_g}(w) \neq \emptyset$. Recall that $\overrightarrow{T}$ is the rooted tree on the same vertex set as $\overrightarrow{V}$ consisting of the arcs of $\overrightarrow{V}$ which were exposed in the course of the breadth-first search beginning at v, with $\overrightarrow{T_w}$ being the subtree of $\overrightarrow{T}$ rooted at w.

If $x \in \overrightarrow{V_g}$, the probability that $\overrightarrow{S_g}(D \setminus V, x) \neq \emptyset$ is at most $(2 + o(1))\varepsilon$ by (6.4). We reuse the term malevolent to mean vertices x such that either $x \in \overrightarrow{V_g}$ and $\overrightarrow{S_g}(D \setminus V, x) \neq \emptyset$ or $x \in \overrightarrow{V_{<g}}$ and x has a descendant y in $\overrightarrow{T}$ such that $y \in \overrightarrow{V_g}$ and $\overrightarrow{S_g}(D \setminus V, y) \neq \emptyset$. Every vertex y in $\overrightarrow{V_g}$ has at most g ancestors in $\overrightarrow{T_g}$, so the expected number of malevolent vertices is at most

$$\mathbb{E}\left[|\overrightarrow{V_g}| \middle| \overrightarrow{V_g} \neq \emptyset\right] (2 + o(1))\varepsilon g = O\left(\frac{\log^3 n\varepsilon^3}{\varepsilon}\right) = o(n\varepsilon^2).$$

Suppose now that w is a benevolent vertex. We let

$$\overrightarrow{U}_w^{(1)} = \overrightarrow{T_w} \cup \bigcup_{x \in \overrightarrow{T_w} \cap \overrightarrow{V_g}} \overrightarrow{S_{\leq g}}(D \setminus V, x).$$

For $m \geq 1$ we let $\overrightarrow{Z}_w^{(m)}$ be the set of vertices x in $\overrightarrow{V}$ such that there is an arc from a vertex in $\overrightarrow{U}_w^{(m)}$ to x, and we let

$$\overrightarrow{U}_w^{(m+1)} = \bigcup_{x \in \overrightarrow{Z}_x^{(m)}} \overrightarrow{U}_x^{(1)} \setminus \bigcup_{i=1}^{m} \overrightarrow{U}_w^{(i)}.$$

For $m \geq 1$, let $\mathcal{B}_w^{(m)}$ be the event that none of the sets $\overrightarrow{U}_w^{(1)}, \ldots \overrightarrow{U}_w^{(m)}$ contains malevolent vertices. Using the same calculations as in (6.16) and (6.17), we get

$$\mathbb{E}\left[|\overrightarrow{U}_w^{(1)}| \middle| v \in B \wedge \mathcal{B}_w^{(1)}\right] = \frac{O\left(\varepsilon^{-3}\log^5 n\varepsilon^3\right)}{|\overrightarrow{V}|}, \tag{6.33}$$

and we can show that

$$\mathbb{E}\left[|\overrightarrow{U}_w^{(m)}| \middle| v \in B \wedge \mathcal{B}_w^{(m)}\right] = \frac{n}{|\overrightarrow{V}|} O\left(\frac{C^m \log^{5m} n\varepsilon^3}{(n\varepsilon^3)^m}\right) \tag{6.34}$$

for a constant C. Suppose now that $\mathcal{B}_w^{(m)}$ holds. The probability that there is an arc from $\overrightarrow{U}_w^{(m)}$ to a malevolent vertex is

$$\frac{1}{|\overrightarrow{V}|} O\left(\frac{C^m \log^{5m} n\varepsilon^3}{(n\varepsilon^3)^m}\right) O\left(\frac{\log^3 n\varepsilon^3}{\varepsilon}\right).$$

The probability that one of the sets $\overrightarrow{U}_w^{(1)}, \overrightarrow{U}_w^{(2)}, \ldots$ contains a malevolent vertex is therefore

$$O\left(\frac{\log^3 n\varepsilon^3}{\varepsilon|\overrightarrow{V}|}\right) O\left(\sum_{m\geq 1} \frac{C^m \log^{5m} n\varepsilon^3}{(n\varepsilon^3)^m}\right) = \frac{1}{|\overrightarrow{V}|} O\left(\frac{\log^8 n\varepsilon^3}{n\varepsilon^4}\right).$$

Thus, the expected number of vertices w in $\overrightarrow{V}$ which has a malevolent descendant is

$$O\left(\frac{\log^8 n\varepsilon^3}{n\varepsilon^4}\right) = o(n\varepsilon^2). \tag{6.35}$$

If we now let $\mathcal{B}_w^{(\infty)}$ be the event that the sets $\overrightarrow{U}_w^{(m)}$ for $m = 1, 2, \ldots$ do not contain any malevolent vertices, and let $\overrightarrow{U_w} = \bigcup_{m\geq 1} \overrightarrow{U}_w^{(m)}$, then by (6.33) and (6.34),

$$\mathbb{E}\left[\overrightarrow{U_w} \middle| v \in B \wedge \mathcal{B}_w^{(\infty)}\right] = \frac{1}{|\overrightarrow{V}|} O\left(\frac{\log^5 n\varepsilon^3}{\varepsilon^3}\right).$$

If $\mathcal{B}_w^{(\infty)}$ holds, the probability that $|\overrightarrow{U_w}| \geq k_0$ is therefore

$$\frac{O\left(\varepsilon^{-3}\log^5 n\varepsilon^3\right)}{|\overrightarrow{V}|k_0} = \frac{1}{|\overrightarrow{V}|} O\left(\frac{\log^4 n\varepsilon^3}{\varepsilon}\right).$$

The expected number of vertices w in $\overrightarrow{V}$ such that none of the sets $\overrightarrow{U}_w^{(1)}, \overrightarrow{U}_w^{(2)}, \ldots$ contains malevolent vertices, and $|\overrightarrow{U_w}| \geq k_0$, is thus

$$O\left(\frac{\log^4 n\varepsilon^3}{\varepsilon}\right) = o(n\varepsilon^2). \tag{6.36}$$

The number of vertices w in $\overrightarrow{V}$ such that $w \in B$ is counted by (6.35) and (6.36), and is $o(n\varepsilon^2)$. By symmetry, the same holds for $\overleftarrow{V}$. This completes the proof of the lemma. □

Proof of Lemma 6.7. Lemmas 6.10 and 6.12 together with Chebyshev's inequality imply that the number Y of vertices v such that $\overrightarrow{S_g}(v) \neq \emptyset$ and $\overleftarrow{S_g}(v) \neq \emptyset$ is a.a.s. $(4 + o(1))\varepsilon^2 n$. To finish the proof of Lemma 6.7, we will show that a.a.s. $|X - Y| < o(\varepsilon^2 n)$.

Let us calculate the expected number of vertices v in $D(n,p)$ such that $\overrightarrow{S_g}(v) \neq \emptyset$, $\overleftarrow{S_g}(v) \neq \emptyset$ and $|\overrightarrow{S_g}(v)| < 1.9\varepsilon n$. From Lemma 6.3 it follows that $\mathbb{P}[k_0 \leq |\overrightarrow{S}(v)| < 1.9\varepsilon n] = o(\varepsilon)$. On the other hand, Lemmas 6.3 and 6.11 imply that the probability that $\overrightarrow{S}_g(v) \neq \emptyset$ and $|\overrightarrow{S}(v)| < 1.9\varepsilon n$ is $o(\varepsilon)$. Furthermore, if we assume that $|\overrightarrow{S}(v)| < 1.9\varepsilon n$, then

$$\begin{aligned}\mathbb{P}\left[\overleftarrow{S_g}(v) \neq \emptyset \middle| |\overrightarrow{S}(v)| < 1.9\varepsilon n\right] &\leq \mathbb{P}\left[|S_g(n,p)| > 0\right] \\ &\quad + \mathbb{P}\left[\exists \text{ an arc from } \overrightarrow{S_g}(v) \text{ to } v\right] \\ &\leq (2 + o(1))\varepsilon + 1.9\varepsilon = (3.9 + o(1))\varepsilon.\end{aligned}$$

Thus the probability that $\overrightarrow{S_g}(v) \neq \emptyset$, $\overleftarrow{S_g}(v) \neq \emptyset$ and $|\overrightarrow{S}(v)| < 1.9\varepsilon n$ is $o(\varepsilon^2)$. By symmetry, the probability that $\overrightarrow{S_g}(v) \neq \emptyset$, $\overleftarrow{S_g}(v) \neq \emptyset$ and $|\overleftarrow{S}(v)| < 1.9\varepsilon n$ is also $o(\varepsilon^2)$. By Markov's inequality, we therefore a.a.s. have $|A \backslash B| = o(\varepsilon^2 n)$.

On the other hand, Lemma 6.11 asserts that the probability that $\overrightarrow{S_g}(v) = \emptyset$ and $|\overrightarrow{S}(v)| \geq 1.9\varepsilon n$ is $o(\varepsilon)$. If this event happens, the probability that $|\overleftarrow{S}(v)| \geq 1.9\varepsilon n$ is still $(2 + o(1))\varepsilon$. In the same way one can argue that the probability that $\overleftarrow{S_g}(v) = \emptyset$, $|\overleftarrow{S}(v)| \geq 1.9\varepsilon n$ and $|\overrightarrow{S}(v)| \geq 1.9\varepsilon n$ is $o(\varepsilon^2)$. Thus, by Markov's inequality $|B \setminus A| = o(\varepsilon^2 n)$ a.a.s. □

Bibliography

[1] Réka Albert, Albert-László Barabási, and Hawoong Jeong. Diameter of the world wide web. *Nature*, 401:130–131, 1999.

[2] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience [John Wiley & Sons], New York, second edition, 2000. ISBN 0-471-37046-0. With an appendix on the life and work of Paul Erdős.

[3] Krishna B. Athreya and Peter E. Ney. *Branching processes*. Die Grundlehren der mathematischen Wissenschaften, Band 196. Springer-Verlag, New York, 1972.

[4] Krystyna T. Balińska, Louis V. Quintas, and Jerzy Szymański. Random recursive forests. In *Proceedings of the Fifth International Seminar on Random Graphs and Probabilistic Methods in Combinatorics and Computer Science (Poznań, 1991)*, volume 5, pages 3–12, 1994.

[5] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. ISSN 0036-8075.

[6] Edward A. Bender and E. Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *J. Combinatorial Theory Ser. A*, 24(3):296–307, 1978.

[7] Béla Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European J. Combin.*, 1(4):311–316, 1980. ISSN 0195-6698.

[8] Béla Bollobás. The evolution of random graphs. *Trans. Amer. Math. Soc.*, 286(1):257–274, 1984. ISSN 0002-9947.

[9] Béla Bollobás. The evolution of sparse graphs. In *Graph theory and combinatorics (Cambridge, 1983)*, pages 35–57. Academic Press, London, 1984.

[10] Béla Bollobás. Random graphs. In *Combinatorics (Swansea, 1981)*, volume 52 of *London Math. Soc. Lecture Note Ser.*, pages 80–102. Cambridge Univ. Press, Cambridge, 1981.

[11] Béla Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001. ISBN 0-521-80920-7; 0-521-79722-5.

[12] Béla Bollobás and Oliver M. Riordan. Mathematical results on scale-free random graphs. In *Handbook of graphs and networks*, pages 1–34. Wiley-VCH, Weinheim, 2003.

[13] Béla Bollobás and Andrew Thomason. Threshold functions. *Combinatorica*, 7(1):35–38, 1987. ISSN 0209-9683.

[14] Amin Coja-Oghlan and Mihyun Kang. The evolution of the min-min random graph process. submitted.

[15] Paul Erdős and Alfréd Rényi. On the strength of connectedness of a random graph. *Acta Math. Acad. Sci. Hungar.*, 12:261–267, 1961. ISSN 0001-5954.

[16] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

[17] Paul Erdős and Alfréd Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959. ISSN 0033-3883.

[18] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Bull. Inst. Internat. Statist.*, 38:343–347, 1961.

[19] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. *Comput. Commun. Rev.*, 29: 251, 1999.

[20] Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. book in preparation. version of October 23, 2006.

[21] Edgar N. Gilbert. Random graphs. *Ann. Math. Statist.*, 30:1141–1144, 1959.

[22] Theodore E. Harris. *The theory of branching processes*. Die Grundlehren der Mathematischen Wissenschaften, Band 119. Springer-Verlag, Berlin, 1963.

[23] Jean Jacod and Albert N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 2003. ISBN 3-540-43932-3.

[24] Svante Janson. Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Process. Appl.*, 110(2): 177–245, 2004. ISSN 0304-4149.

[25] Svante Janson. Asymptotic degree distribution in random recursive trees. *Random Structures Algorithms*, 26(1-2):69–83, 2005. ISSN 1042-9832.

[26] Svante Janson and Joel Spencer. A point process describing the component sizes in the critical window of the random graph evolution. *Combin. Probab. Comput.*, 16(4):631–658, 2007.

[27] Svante Janson, Tomasz Łuczak, and Andrzej Ruciński. *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000. ISBN 0-471-17541-2.

[28] Mihyun Kang and Taral Guldahl Seierstad. The critical phase for random graphs with a given degree sequence. to appear in *Combin. Probab. Comput.*, .

[29] Mihyun Kang and Taral Guldahl Seierstad. The phase transition of the minimum degree random multi-graph process. to appear in *Random Structures Algorithms*, .

[30] Mihyun Kang, Youngmee Koh, Sangwook Ree, and Tomasz Łuczak. The connectivity threshold for the min-degree random graph process. *Random Structures Algorithms*, 29(1):105–120, 2006. ISSN 1042-9832.

[31] Richard M. Karp. The transitive closure of a random digraph. *Random Structures Algorithms*, 1(1):73–93, 1990. ISSN 1042-9832.

[32] Richard M. Karp and Michael Sipser. Maximum matchings in sparse random graphs. In *Proceedings of the Twenty-Second Annual IEEE Symposium on Foundations of Computing*, pages 364–375, 1981.

[33] János Komlós and Endre Szemerédi. Limit distribution for the existence of Hamiltonian cycles in a random graph. *Discrete Math.*, 43(1):55–63, 1983. ISSN 0012-365X.

[34] Malwina Luczak and Tomasz Łuczak. The phase transition in the cluster-scaled model of a random graph. *Random Structures Algorithms*, 28(2):215–246, 2006. ISSN 1042-9832.

[35] Tomasz Łuczak. Component behavior near the critical point of the random graph process. *Random Structures Algorithms*, 1(3):287–310, 1990. ISSN 1042-9832.

[36] Tomasz Łuczak. On the equivalence of two basic models of random graphs. In *Random graphs '87 (Poznań, 1987)*, pages 151–157. Wiley, Chichester, 1990.

[37] Tomasz Łuczak. Random trees and random graphs. In *Proceedings of the Eighth International Conference "Random Structures and Algorithms" (Poznan, 1997)*, volume 13, pages 485–500, 1998.

[38] Tomasz Łuczak. The phase transition in the evolution of random digraphs. *J. Graph Theory*, 14(2):217–223, 1990. ISSN 0364-9024.

[39] Tomasz Łuczak and Joel E. Cohen. Giant components in three-parameter random directed graphs. *Adv. in Appl. Probab.*, 24(4):845–857, 1992. ISSN 0001-8678.

[40] Tomasz Łuczak and Taral Guldahl Seierstad. The critical behavior of random digraphs. submitted.

[41] Hosam M. Mahmoud and R. T. Smythe. Asymptotic joint normality of outdegrees of nodes in random recursive trees. *Random Structures Algorithms*, 3(3):255–266, 1992. ISSN 1042-9832.

[42] Brendan D. McKay. Asymptotics for symmetric 0-1 matrices with prescribed row sums. *Ars Combin.*, 19(A):15–25, 1985. ISSN 0381-7032.

[43] Michael Molloy and Bruce Reed. The size of the giant component of a random graph with a given degree sequence. *Combin. Probab. Comput.*, 7(3):295–305, 1998. ISSN 0963-5483.

[44] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. In *Proceedings of the Sixth International Seminar on Random Graphs and Probabilistic Methods in Combinatorics and Computer Science, "Random Graphs '93" (Poznań, 1993)*, volume 6, pages 161–179, 1995.

[45] Mark E. J. Newman, Steven H. Strogatz, and Duncan J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, Jul 2001. doi: 10.1103/PhysRevE.64.026118.

[46] Robert W. Robinson and Nicholas C. Wormald. Almost all cubic graphs are Hamiltonian. *Random Structures Algorithms*, 3(2):117–125, 1992. ISSN 1042-9832.

[47] Robert W. Robinson and Nicholas C. Wormald. Almost all regular graphs are Hamiltonian. *Random Structures Algorithms*, 5(2):363–374, 1994. ISSN 1042-9832.

[48] Andrzej Ruciński and Nicholas C. Wormald. Random graph processes with degree restrictions. *Combin. Probab. Comput.*, 1(2):169–180, 1992. ISSN 0963-5483.

[49] Andrzej Ruciński and Nicholas C. Wormald. Connectedness of graphs generated by a random d-process. *J. Aust. Math. Soc.*, 72(1):67–85, 2002. ISSN 1446-7887.

[50] Neil J. A. Sloane. *The On-Line Encyclopedia of Integer Sequences*, 2006. URL `www.research.att.com/~njas/sequences/`.

[51] Robert T. Smythe and Hosam M. Mahmoud. A survey of recursive trees. *Theory Probab Math Statist*, 51:1–27, 1995.

[52] Richard P. Stanley. *Enumerative combinatorics. Vol. 1*, volume 49 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1997. ISBN 0-521-55309-1; 0-521-66351-2. With a foreword by Gian-Carlo Rota, Corrected reprint of the 1986 original.

[53] András Telcs, Nicholas C. Wormald, and Sanming Zhou. Hamiltonicity of random graphs produced by 2-processes. to appear in *Random Structures Algorithms*.

[54] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[55] Herbert S. Wilf. *generatingfunctionology*. A K Peters Ltd., Wellesley, MA, third edition, 2006. ISBN 978-1-56881-279-3; 1-56881-279-5.

[56] Nicholas C. Wormald. Differential equations for random processes and random graphs. *Ann. Appl. Probab.*, 5(4):1217–1235, 1995. ISSN 1050-5164.

[57] Nicholas C. Wormald. The differential equation method for random graph processes and greedy algorithms. In M. Karoński and H-J. Prömel, editor, *Lectures on approximation and randomized algorithms.*, pages 75–152. PWN, Warsaw, 1999.

[58] Nicholas C. Wormald. *Some Problems in the Enumeration of Labelled Graphs.* PhD thesis, Newcastle University, 1978.

Erklärung

Hiermit erkläre ich, dass

- ich die vorliegende Dissertationsschrift selbstständig ohne fremde Hilfe verfasst und nur die angegebene Literatur und Hilfsmittel verwendet habe,
- ich mich nicht bereits anderwärts um einen Doktorgrad beworben habe oder einen solchen besitze, und
- mir die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät II der Humboldt-Universität zu Berlin vom 17.1.2005, zuletzt geändert am 13.02.2006, bekannt ist. (Amt. Mitteilungsblatt Nr. 34/2006)

Berlin, den 29. Januar 2007

Taral Guldahl Seierstad